Computational analysis of gene content in Xenacoelomorpha

Bartlomiej Piotr Tomiczek

Department of Genetics Evolution and Environment

University College London

Supervisors:
Max Telford, Christophe Dessimoz

Submitted for Degree of Doctor of Philosophy

Declaration

I, Bartlomiej Piotr Tomiczek, confirm that work presented in this thesis is my own. Where information has been derived from other sources, I confirm this has been indicated in the thesis

Signed:

Acknowledgments

*I would like to thank my supervisor **Max Telford** for being for being not only my tutor but also a great support for the last few years. Thank you for all the good encouragement and care.*

*Thanks to **Christophe Dessimoz** for a kind word and valuable advice. To **Adrian Altenhoff** for help with OMA standalone software. To **Albert Poustka** for providing the genome data. To **Hervé Philippe** for help with the phylogenetic analysis. To **Jean-François** for help with the data quality analysis. To **Jeremy Levy** for help with OrthoMCL. To **Nives Škunca** for GO analysis. To Steven Müller the help with the HOG analysis.*

*Many thanks to **Bernhard Egger** for giving me a start in taxonomy and teaching me animal collection techniques. Many thanks to my roommates, **Mario Dos Reis** for teaching me basics of molecular evolution and to **Francois Lapraz** for shearing drawing useful graphs on the walls of GEE. To **Johannes Girstmair** and **Ania Czarkwiani** for always putting smile on my face. To **David Dylus** and **Maria** for keeping me busy during the nights.*

*Thanks to **Ziheng Yang, Konstantinos Angelis, Jose Antonio Barba Montoya** and **Daniel Dalquen** for all the good advice and even better company in the office. To **Paola Oliveri** and **Libero Petrone** for cultural education. To **Lucas Daniel Wittwer**, **Ivana Piližota** and other members of Dessimoz lab for storming discussions on group meetings.*

And most importantly to my wife and daughter for the support and motivation. To my parents for having faith in me through all these years and for helping me in difficult moments.

Table of Contents

Computational analysis of gene content in Xenacoelomorpha

Abstract **1**

Chapter 1:

1.1: Introduction **2**

1.1.1 Xenacoelomorpha - Morphology and habitat 2

1.1.2 The relationship between Acoela, Nemertodermatida and Xenoturbella 3

1.1.2.1 Xenoturbellida 5

1.1.2.2 Acoela 7

1.1.2.2.1 Phylogenetic relationships within the Acoela 11

1.1.2.3 Nemertodermatida 13

1.1.3 Synapomorphies within Xenacoelomorpha 14

1.1.4 Phylogenetic position of Xenacoelomorpha inferred from their morphology 14

1.1.5 The hypothesis about evolution of Bilateria 15

1.1.6 The history of molecular phylogenetic studies of Xenacoelomorpha 16

1.1.7 Evolutionary implications of phylogenetic positions of Xenacoelomorpha at the base of all Bilateria. 22

1.1.8 Evolutionary implications of phylogenetic positions of Xenacoelomorpha as a sister group to Ambulacraria. 23

1.1.9 Outline 24

Chapter 2:

Quality assessment of the Xenacoelomorpha genomic and transcriptomic sequences **26**

2.1 Introduction **26**

2.1.1 Potential factors influencing the quality of high-throughput sequencing data 29

2.1.2 Testing the quality of Xencoelomorpha Next Generation Sequencing data 33

2.1.3 The aims of the quality assessment 33

2.2 Materials and methods 37

2.2.1 Animal collection, DNA/mRNA extraction and sequencing 37

2.2.2 Read processing and the genome complexity analysis 37

2.2.3 Genome size estimation 37

2.2.4 The Genome Assembly and assembly properties 37

2.2.5 The Transcriptome assembly 38

2.2.6 Assembly decontamination 38

2.3 Results 40

2.3.1 The Xenacoelomorpha genomes are difficult to assemble due to heterozygosity and high repeat content 40

2.3.2 Genome size estimation 44

2.3.3 The assembly quality 48

2.3.3.1 The N50 48

2.3.3.2 Number of contigs greater than 10kb 49

2.3.3.3 The N50 scaffold length 49

2.3.5 The presence of core Eukaryote proteins in animal proteomes 58

2.3.5.1 The use of multiple proteomes improves the ability to find core Eukaryote genes within the Xenacoelomorpha clade 59

2.3.6 Detecting contamination in the predicted protein datasets 61

2.4 Conclusions 63

Chapter 3:

Analysis of gene family content in Xenacoelomorpha genomes using PhylomeDB database . 65

3.1 Introduction 65

3.1.1 Gene families as clade specific characteristics 66

3.1.2 PhylomeDB database as a resource in gene family search 66

3.2 Materials and methods 68

3.2.1 Identifying clade specific genes/gene families' candidates using PhylomeDB databse 68

3.2.2 Verifying absence of putative clade specific genes/gene families using BLAST searching of online databases 70

3.2.3 Xenacoelomorph sequence data 71

3.2.4 Testing for presence of a family members using family-RBH (Reciprocal Best Hit) optimising an appropriate e-value cutoff and p-value tolerance 73

3.3 Results 75

3.3.1 Preliminary identification of Bilaterian, Deuterostome and Chordate specific gene families 75

3.3.2 Verifying absence of clade specific gene in outgroup taxa using NCBI database 75

3.3.3 Bilaterian specific gene families in Xenacoelomorpha 75

3.3.4 Deuterostome specific gene families in Xenacoelomorpha 77

3.3.5 Apparent chordate specific gene families also present in Ambulacraria 78

3.3.6 An implication for the evolution of Xenacoelomorpha 81

3.3.8 Correlation of the gene family loss with morphological complexity 82

3.4 Conclusions 85

Chapter 4:

Impact of automated orthology group assignment on the reconstruction of lophotrochozoan phylogeny 87

4.1 Introduction 87

4.2 Materials and methods 93

4.2.1 Transcriptome assembly and peptide prediction 93

4.2.2 Sequence Processing 93

4.2.3 CEGMA pipeline 94

4.2.4 OMA pipeline 94

4.2.5 OrthoMCL pipeline 94

4.2.7 Protein sequence alignments and Phylogenetic Analyses 95

4.3 Results 96

4.3.1 Comparison of orthology groups' sizes inferred with OMA, CEGMA and OrthoMCL pipelines 96

4.3.2 Supermatrix density comparison 97

4.3.3 Consistency with current taxonomy 98

4.3.4 Lophotrochozoa phylogeny inference 101

4.3.4.1 The analysis of CEGMA dataset 106

4.3.4.2 The analysis of OrthoMCL dataset 109

4.4 Conclusions 112

Chapter 5:

The construction of Metazoa gene family database, involving protein sets from 67 species, including 8 Xenacoelomorpha 116

5.1 Introduction 116

5.1.1 Construction of gene family database 117

5.1.2 Inferring the orthology relations between proteins using the OMA standalone package . 119

5.1.4 Accessing hierarchical group content on different taxonomic levels – familyanalizer.py . 120

5.2 Materials and methods 123

5.3 Results **126**

5.3.1 Metazoa gene family database 126

5.3.1.1 Basic characteristics of our gene family database 127

5.3.2 Ancestral Metazoa gene families in our OMA standalone database 127

5.3.3 Metazoa gene family database including 4 new Xenacoelomorpha proteomes 128

5.4 Conclusions **147**

Chapter 6:

Phylogenetic analysis of Xenacoelomorpha based on orthology groups created using whole genomic sequences from 67 Metazoa species. **148**

6. 1 Introduction **146**

6.2 Results **152**

6.3 Final Conclusions **162**

Table of Contents – Figures

Chapter 1:

Page 4

Figure 1.1. Phylogenetic relations within Xenacoelomorpha, where Nemertodermatida are sister group to Acoela, and Xenoturbelida (*Xenoturbella*) are basal within Xenacoelomorpha.

Page 6

Figure 1.2. b) Scanning electron micrograph of adult *Xenoturbella bocki*, the specimen, seen in dorsal view, has contracted lengthwise into a rugby ball shape but the radial and longitudinal grooves are visible (photo courtesy of Ake Franzen and Bjorn Afzelius). Scale bar 0.1 cm. d) Dorsal photograph of adult *Xenoturbella bocki*. Note the lightening of pigment at an anterior end (A, anterior; P, posterior). Scale bar 0.1 cm.

Page 7

Figure 1.3. Phylogeny of *Xenoturbella* based on mitochondrial DNA sequences (15,532 base-pair alignment) inferred with Maximum Likelihood method with GTR + Γ (General Time Reversible model with Gamma approximation for rate variation among sites)

Page 8

Figure 1.4. The photography of the pink pigmented *Paratomellarubra*from Filey coast in north Yorkshire. (Picture Egger and Tomiczek).

Page 10

Figure 1.5. Images of sensory structures of live *Symsagittiferaroscoffensis* analysed in this study.

Figure 1.6. Image of a mature and live specimen of *Isodiametrapulchra* analysed in this study without (left) and with superimposed colors (right) to illustrate the general morphology of acoels.

Page 12

Figure 1.7. Cladogram of the Acoelomorpha with partial family-level systematics of the Acoela.

Page 13

Figure 1.8. Gravid adult of *Meara stichopi* collected in throughout the winter between 2009/2010 - 2013/2014 at collection sites around Bergen, Norway by AinaBørve .

Page 17

Figure 1.9. Three scenarios for the phylogenetic position of Acoela, Nemertodermatida and Xenoturbellida.

Chapter 2:

Page 27

Figure 2.1. The block scheme representing the workflow of data processing, with the quality tests applied on read level (1), genome assembly level (2), genomic gene prediction level (3a), transcriptomic gene prediction level (3b) and completed protein set for each species (3c).

Page 41

Figure 2.2. A High rate of the heterozygous variation in the Xenacoelomorpha genomes in comparison to the reference genome assemblies.

Page 43

Figure 2.4. A High rate of sequence repeats in the Xenacoelomorpha genomes.

Page 44

Figure 2.4. A histogram of 51-mer frequencies for each set of Ilumina paired reads from Xenacoelomorpha genome sequencing.

Page 46

Figure 2.5. Estimated genome size of Xenacoelomorpha species.

Table 2.1 Estimated genome coverage calculated based on genome size estimation and read number.

Page 48

Figure 2.6. N50 contig size of the Xenacoelomorpha assemblies. N50 metric shows low contiguity of *Symsagittiferaroscoffensis, Meara stichopi, Nemertodermawestbladi, Pseudophanostomavariabilis*and *Paratomellarubra*.

Page 48

Figure 2.7. *Xenoturbella* and *Symsagetifera*assemblies contain the most contigs greater than 10kb.

Page 50

Figure 2.8. N50 scaffold length of the Xenacoelomorpha assemblies.

Page 51

Figure 2.9. *Meara stichopi*and the *Pseudophanostomavariabilis*assemblies contain the highest percentage of gaps in scaffold.

Page 53

Figure 2.10. Number of predicted ORFs from the Trinity transcriptome assemblies.

Page 55

Figure 2.11. Number of predicted genes from the genome assemblies of Xenacoelomorpha.

Page 57

Figure 2.12. Number of predicted genes after clustering with CD-HIT.

Table 2.1 Number of predicted ORFs for each of the sequenced Xenacoelomorpha species.

Page 58

Figure 2.13. The presence of the core Eukaryotic proteins in animal proteomes.

Page 58

Figure 2.14. Proportion of the present core Eukaryote gene in the subset of the Xenacoelomorpha proteomes improves with the number of proteomes used in the analysis.

Page 61

Figure 2.15. Proportion of present frequently present Eukaryote gene (present in at least 50% of taxa) in the subset of the Xenacoelomorpha proteomes improves with the number of proteomes used in the analysis.

Page 63

Figure 2.16. Proportion of contaminated sequences coming from human (blue), and all sorces identified by Kraken sofrware in Xenacoelomorpha protein predictions.

Chapter 3

Page 70

Figure 3.1. A schematic representation of Chordata, Deuterostomia, Bilateria clade specific gene families.

Page 75

Figure 3.2. The estimation of the E-value threshold and the p-value tolerance coefficient parameters in the family-RDH algorithm.

Page 78

Figure 3.3. Bilaterian specific gene families are present in in Xenacoelomorpha proteomes.

Page 80

Figure 3.4. Deterostome specific gene families are present in Xenacoelomorpha proteomes.

Page 82

Figure 3.5. Apparent chordate specific gene families also present in Ambulacraria proteomes.

Page 85

Figure 3.6. High proportion of Metazoa ancestral gene families is present in extant animals.

Page 85

Figure 3.7. High proportion of Bilateria ancestral gene families is present in extant animals.

Chapter 4

Page 89

Figure 4.1. The flowchart of the OMA, CEGMA and OrthoMCL phylogenetic pipelines.

Page 92

Table 4.1 Next Generation Sequencing data for 30 species used in the analysis. The dataset includes 30 most complete sets of proteins (proteomes) from 19 lophotrochozoans, 4 deuterostomes, 4 ecdysozoans and 3 non-bilaterians.

Page 92

Figure 4.2. Phylogenetic tree inference based on a groups of orthologous (blue) and paralogous sequences (red).

Page 99

Figure 4.3 Distribution of group sizes.

Page 100

Figure 4.4. Histogram represents the number of amino acid positions with different supermatrix density.

Page 102

Figure 4.5. OMA orthology groups tend to recover the monophyly of Lophotrochozoa more frequently.

Page 103

Figure 4.6. The example of gene trees, in which Lophotrochozoa are monophyletic on a gene tree calculated based on OMA orthology group, but not in the corresponding CEGMA orthology group.

Page 107

Figure 4.7. The Bayesian phylogeny calculated with OMA pipeline using CAT GTR G4 model in PhyloBayes.

Page 108

Figure 4.8. The ML phylogeny calculated with OMA pipeline in RaXML.

Page 110

Figure 4.9. The Bayesian phylogeny calculated with CEGMA pipeline using CAT GTR G4 model with PhyloBayes.

Page 110

Figure 4.10. The ML phylogeny calculated with CEGMA pipeline in RaXML.

Page 112

Figure 4.11. The Bayesian phylogeny calculated with OrthoMCL pipeline using CAT GTR G4 model with PhyloBayes.

Page 113

Figure 4.12. The ML phylogeny calculated with OrthoMCL pipeline in RaXML.

Chapter 5

Page 125

Figure 5.1 The dendrogram representing the phylogenetic relation between 30 species in Metazoa gene family database (dataset_1) used as a leading phylogeny to infer paralogy/ orthologyrelation within gene families.

Page 126

Figure 5.2 The distribution of average family size for 30,932 families calculated for dataset 1.

Page 127

Figure 5.3. The distribution of average family size for 13,878 ancestral to Metazoa (present in Metazoa Last Common Ancestor) families calculated for dataset_1.

Figure 5.4 Histogram representing the distribution of evolutionary distance between members of 13,250 Xenacoelomorpha families.

Page 130

Figure 5.5 The influence of taxa selecton on number of core gene families in the database.

Page 130

Figure 5.6 The influence of taxa selecton on number of Xenacoleomorpha families in the database.

Page 131

Figure 5.7 Three different positions of Xenacoelomorpha on a tree of life used as a leading phylogeny in the inference of gene family database (dataset_2).

Page 132

Figure 5.8 Large Xenacoelomorpha families with 10 or more members tend to overlaping gene content more frequently with two different leading phylogenies.

Page 134

Figures 5.9 Only ancestral to Metazoa families are independent from the phylogenetic position of Xenacoelomorpha (highlighted in green).

Page 136

Figure 5.10 Gene family evolution across Metazoa.

Page 139

Figure 5.11 Xenacoelomorpha are sister group to Nephrozoa at the base of Bilateria with high evolutionary rate on a branch leading to Xenacoelomorpha Last Common Ancestor (0.6 event per gene).

Page 140

Figure 5.12 Xenacoelomorpha are sister group to Ambulacraria (scenario A) with a similar evolutionary rate as other main clades of Metazoa.

Figure 5.13 Xenacoelomorpha didn't lose more genes then other main anima clades, while maintaining low duplication rate.

Page 142

Figure 5.14 Xenacoelomorpha lost more ancestral gene familiessimultaneously with deuterostomes then protostomes.

Page 142

Figure 5.15 Xenacoelomorpha lost more genes simultaneously with Ambulacraria and Ecdysozoa then Chordata and Lophotrochozoa.

Page 144

Figure 5.16 Xenacoelomorpha ancestor is more similar to Xenambulacraria ancestor then to Bilateria ancestor.

Page 145

Figure 5.17 Gene family evolution across Metazoa as inferred from 67 proteomes, Xenacoelomorpha are basal Bilateria (scenario B)).

Page 146

Figure 5.18 Gene family evolution across Metazoa as inferred from 67 proteomes, Xenacoelomorpha are sister group to Ambulacraria (scenario A)).

Chapter 6

Page 152

Figure 6.1. Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 model from full 350,090 Amino Acid positions alignment.

Page 153

Figure 6.2 The jackknife analysis of 100 datasets of 20,000 amino acids each, inferred using CAT+GTR+G4 model in PhyloBayes.

Page 156

Figure 6.3 Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 model based on the full 350,090 amino acid positions alignment, without Acoelomorpha.

Page 157

Figure 6.4 The jackknife analysis of 100 subsets of 20,000 amino acids each, produced using the PhyloBayes CAT+GTR+G4 model, without Acoelomorpha.

Page 159

Figure 6.5 Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 on 31,000 amino acid positions superalignment of genes that reconsolidate monophyletic Ambulacraria.

Page 160

Figure 6.6 Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 on 31,000 amino acid positions superalignment of genes that reconsolidate monophyletic Protostomia.

Page 164

Figure 6.7 Evolution of characteristics within Metazoa according to the scenario, where deuterostomes are not monophyletic (supported by molecular phylogeny with the Site-Heterogeneous CAT+GTR+G4 (Figure 6.1; 6.3; 6.4; 6.6)).

Computational analysis of gene content in Xenacoelomorpha

Bartlomiej Tomiczek

Supervisors: Max Telford, Christophe Dessimoz

Abstract

Xenacoelomorpha are simple, marine worms with net-like nervous systems, no circulatory or respiratory systems and a blind gut. The phylogenetic position of Xenacoelomorpha is the subject of ongoing debate in the literature. The two possible locations for the Xenacoelomorpha within the animal tree are i) as the sister clade to all other bilaterians and ii) as deuterostomes, closely related to the Ambulacraria (echinoderms and hemichordates). The understanding of the phylogenetic position of Xenacoelomorpha has major implications in understanding the appearance of the Bilateria last common ancestor and the direction of the evolutionary process within the animal kingdom. If Xenacoelomorpha are in fact basal bilaterians, they can resemble many similarities to simple acoel-like bilaterian ancestor. However, if Xenacoelomorpha are sister group to Ambulacraria, they likely secondary simplified from a complex, segmented, coelomate Bilateria ancestor.

I analysed the quality of 6 new xenacoelomorph genomic and 7 transcriptomic data sets (*Symsagittifera roscoffensis, Pseudophanostoma variabilis, Paratomella rubra* and *Praesagittifera naikaiensis*, the nemertodermatids *Meara stichopi* and *Nemertoderma westbladi* and the xenoturbellid *Xenoturbella bocki*), and have constructed comprehensive datasets of xenacoelomorph proteins (proteomes (entire set of proteins expressed by a specific organism (UniProt Consortium, 2010))). I used these, together with proteomes from 60 other species, to construct a database of gene families, which have descended from the same common ancestor within the broad range of 67 species within the animal kingdom. Based on inferred orthology/paralogy relations within

these families, I reconstructed the duplications, gains and losses of genes across the Metazoa. The analysis of ancestral gene family content is suggestive for the phylogenetic position of the Xenacoelomorpha, as ancestral Xenacoelomorpha gene content is more similar to inferred Xenambulacraria gene content then to ancestral Bilateria gene content. Moreover, Xenacoelomorpha show more simultaneous gene losses with Ambulacraria then with other major Bilateria clades.

To reconstruct a molecular phylogenetic tree of Xenacoelomorpha, I first established a bioinformatics pipeline for large-scale molecular phylogeny reconstruction, by comparing 3 commonly used automated methods for orthology and paralogy prediction (OMA, CEGMA, OrthoMCL). I tested the application of these methods in constructing phylogenetic matrices from high throughput sequencing data. I used the best performing pipeline to infer the species tree involving 8 Xenacoelomorpha species. Our phylogenetic analysis tentatively supports the placement of Xenacoelomorpha as a sister group of Ambulacraria.

Chapter 1

1.1 Introduction

In this thesis I try to address the ongoing controversy in the literature about the early evolution of Bilateria and the debated phylogenetic position of Xenacoelomorpha. To do that, I analysed the new genomic and transcriptomic resources of xenacoelomorphs. For the analysis we chose represents of each Xenacoelomorpha subphylum, the xenoturbellid *Xenoturbella bocki,* the nemertodermatids *Meara stichopi* and *Nemertoderma westbladi* and the higher acoels *Symsagittifera roscoffensis, Pseudophanostoma variabilis,* and *Praesagittifera naikaiensis* and collected in our lab basally branching *Paratomella rubra.* Here, I will first breathily characterize the current state of knowledge about xenacoelomorphs, and describe their key morphological features, as well as their possible relation with

other species and the history of their phylogenetic position. I discuss the possible placement of Xenacoelomorpha on the tree of life, and possible way of their simplification during evolution.

1.1.1 Xenacoelomorpha - Morphology and habitat

Xenacoelomorpha are worms that are found in marine habitats. Various species of Xenacoelomorpha live dwelling in a mud, or between grains of sand, at the sea bottom or on beach coasts. Ciliary gliding is their primary movement mechanism. Xenacoelomorpha can be divided into three subphyla, Acoelomorpha, Nemertodermatida and Xenoturbellida. All the subphyla do not have a stomatogastric system and a through gut, there is no circulatory, respiratory systems or centralized nervous system. In Acoela the mouth opens directly into the mesoderm, in Nemertodermatida and Xenoturbellida, unciliated cells cover the sack-like gut. The nervous system is located right under the epidermis and they do not have a brain. In Xenoturbellida it is constituted by a simple nerve net without any special concentration of neurons, while in Acoela and Nemertodermatida it is arranged in a series of longitudinal bundles. The statocyst, which is responsible for the balance and the sense of gravity, is the main sensory organ of Xenacoelomorpha. The statocyst is located in anterior part of the body, and is surrounded by nerves. The epidermis is ciliated, which are composed from set of 9 pairs of peripheral microtubules and one or two central microtubules. Unusually, the microtubule pairs from 4 to 7 terminate before the tip to create a shelf like structure.

1.1.2 The relationship between Acoela, Nemertodermatida and Xenoturbella

Xenacoelomorpha can be divided into two sister groups Acoelomorpha and Xenoturbellida, where Acoelomorpha consists of Acoela and Nemertodermatida (see Figure 1.1). Nemertodermatida were originally considered as part of Acoela, but the morphological differences helped to recognize them as a separate clade (Steinböck et al. 1930). The first studies of nucleotide sequences showed that Acoela and Nemertodermatida are paraphyletic, with Nemertodermatida as a sister group to Nephrozoa, and

Acoela as at the base of Bilateria (Jondelius et al. 2002; Wallberg et al. 2007; Paps et al. 2009). However, later studies confirmed the previously recognized monophyly of Acoelomorpha (Westheide and Rieger 2007; Hejnol et al. 2009; Philippe et al. 2007, 2011). *Xenoturbella bocki*, until recently the lonely member of Xenoturbellida (Rouse et al. 2016), was first grouped together with Acoelomorpha as part of Platyhelminthes (Westblad 1949; Franzen et al. 1987), but several molecular phylogenetic analysis of nucleotide data have credibly shown that neither group is closely, to this phylum (Wallberg et al. 2007; Jondelius et al. 2002; Littlewood et al. 2001; Ruiz-Trillo et al. 2002; Telford et al. 2000, 2003; Philippe et al. 2007). More recently, it has been generally accepted that the Acoelomorpha and Xenoturbellida are each other's closest relatives, based on the phylogenetic analysis of molecular data (collectively the Xenacoelomorpha) (Hejnol et al. 2009; Philippe et al. 2011) (see Figure 1.1). The analysis of myosin heavy chain II sequences (Ruiz-Trillo et al. 2002), Hox genes (Fritsch et al. 2008) and the mitochondrial genome (Ruiz-Trillo et al. 2004) supports the monophyly of the Xenacoelomorpha. However, the position of the group within animal kingdom is less clear. Two possible locations of the Xenacoelomorpha clade on an animal phylogeny are considered, i) as the sister clade to all other bilaterians (Hejnol et al. 2009; Cannon et al. 2016; Rouse et al. 2016) and ii) as deuterostomes, most closely related to the Ambulacraria (echinoderms and hemichordates) (Philippe et al. 2011). In this thesis, I have analysed the sequences of 7 new Xenacoelomorpha genomes and perform the phylogenetic analysis of orthologous groups from 67 animal species to gather more evidence on the phylogenetic position of Xenacoelomorpha within animal kingdom. I shortly characterised the key morphological features of each of the subphyla, which Xenacoelomorpha consists of, and described the relations within the Xenoturbellida, Acoela and Nemertodermatida clades.

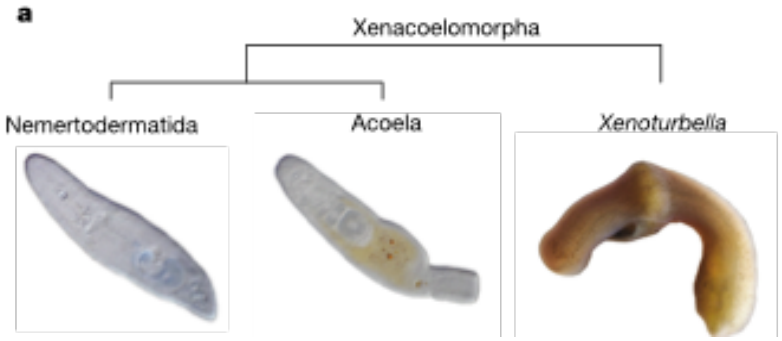


Figure 1.1. Phylogenetic relations within Xenacoelomorpha, where Nemertodermatida are sister group to Acoela, and Xenoturbelida (*Xenoturbella*) are basal within Xenacoelomorpha (Cannon et al. 2016).

1.1.2.1 Xenoturbellida

Until recently, *Xenoturbella bocki*, which genomic sequence I analysed in this thesis, was the only known member of Xenoturbellida (Rouse et al. 2016). This simple droplet shaped worm has been a subject of great debate in the literature, because of the difficulties concerning the placing within the animal kingdom. The specimens sequenced here were found at the Swedish coast in Kristineberg Marine Research Station by joint efforts of Max Telford and Albert Poustka Lab, sequenced and assembled thanks to Xenacoelomorpha Genome Project 2014. *Xenoturbella bocki* is a small (typically 2 cm long), yellowish-brown, flattened worm first found in 1915 by Sixten Bock on the West coast of Sweden and first described in 1949 by Westblad. *Xenoturbella's* body exhibits two furrows, circumferential and horizontal lateral furrow, which coincide with the increased thickness of the nerve net, and lightening of pigmentation around the anterior end (see Figure 1.2). Two other features are visible, the main sensory organ, the statocyst, is present at the anterior part of the body, and mid-ventral mouth opening, which opens into blind gut cavity. There is no stomatogastric system, no anus, no circulatory, respiratory systems or brain.

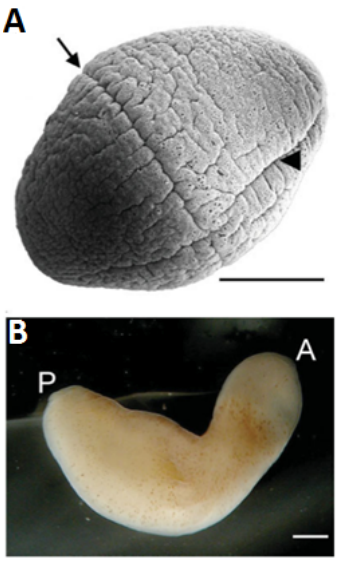


Figure 1.2. A) Scanning electron micrograph of adult *Xenoturbella bocki*, the specimen, seen in dorsal view, has contracted lengthwise into a rugby ball shape but the radial and longitudinal grooves are visible (photo courtesy of Ake Franzen and Bjorn Afzelius). Scale bar 0.1 cm. B) Dorsal photograph of adult *Xenoturbella bocki*. Note the lightening of pigment at an anterior end (A, anterior; P, posterior). Scale bar 0.1 cm. (Telford et al. 2008).

Recently the knowledge about Xenoturbellida diversity within the clade has improved, thanks to four new species of *Xenoturbella* have been found in depths of the Pacific Ocean (Rouse et al. 2016). The new *Xenoturbella* species are beautifully coloured, ranging from brown to orange to pink to purple, and three of the new species are giants compared to *Xenoturbella bocki* with the largest measuring over 20 cm long (see Figure 1.3). The phylogenetic analysis of whole mitochondrial genomes places the three larger and deep-water species together, as a sister clade to small shallow water species, both from California and Swedish coast.

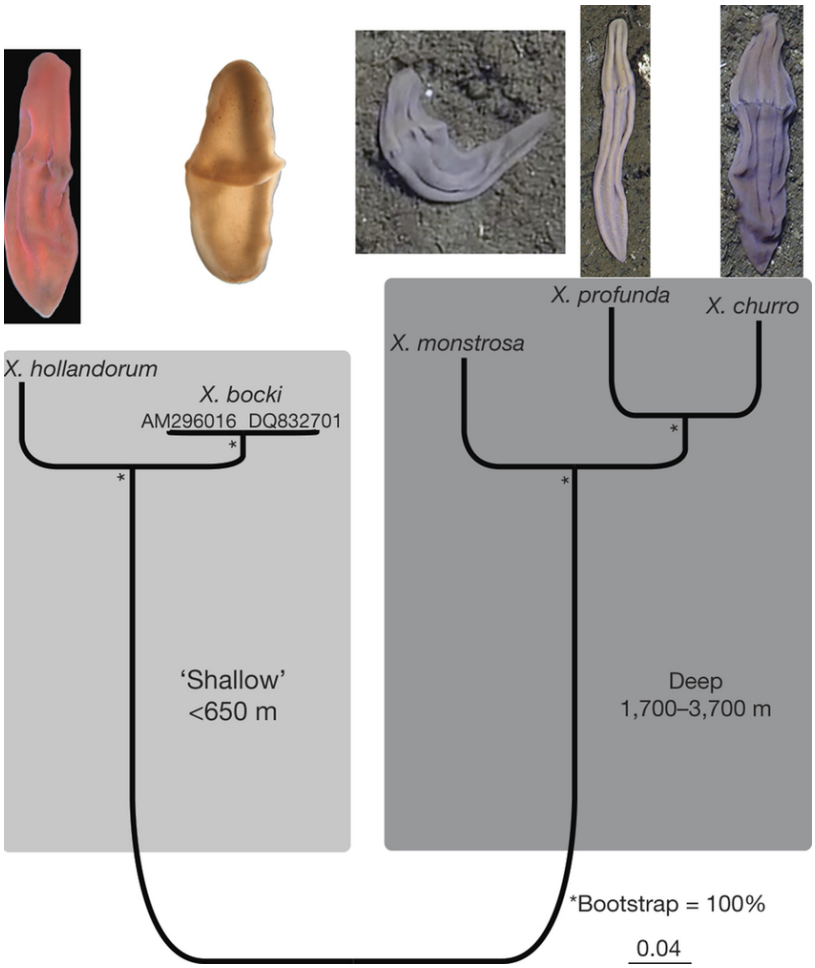


Figure 1.3. Phylogeny of *Xenoturbella* based on mitochondrial DNA sequences (15,532 base-pair alignment) inferred with Maximum Likelihood method with GTR+Γ (General Time Reversible model with Gamma approximation for rate variation among sites) (Rouse et al. 2016).

1.1.2.2 Acoela

Other important members of Xenacoelomorpha clade, whose genomes have been analysed in this thesis, are Acoelomorpha, animals that have been regarded as representatives of early stage in evolution of Bilateria (Gaff 1904; Hyman 1959). Acoelomorpha consists of two main clades, Acoela and Nemertodermatida. One of the acoels, *Paratomella rubra,* I collected, together with Bernhard Egger from the sand of beaches in Filey and Robin Hood's Bay, on the coast of North Yorkshire. The acoels are typically microscopic worms, mostly free living in marine habitats. The nemertodermatid *Meara stichopi* is

an endosymbiont in the digestive system of the sea cucumber, *Stichopus* (Jennings et al. 1971). The diet of the acoels is diverse and ranges from bacteria, algae to crustaceans (shrimps); some eat other worms including members of the same phylum. Most of the acoels are transparent or opaque, but some can be naturally pigmented or by algal symbionts, or by glandular secretions called rhabdoids (see Figure 1.5a). *Paratomella rubra* has a pink pigmentation and is easy to distinguish on white background from other acoels under the binocular microscope (see Figure 1.4). Acoels' body shape ranges from long to droplet shape and can be flat or slender depending on the environment they move in. The epidermis is ciliated and with a characteristic shelf at the tip of the microtubules (see above) and the distinctive rootlet system typical for the phylum (Rieger et al. 1991). Gland cells are positioned in the epidermis and are thought to help with the ciliary motion, but may also act as a defensive mechanism (Pedersen 2006). Besides ciliary gliding, acoels use muscles to move. Dorso-ventral muscles serve to flatten the body, and the musculature of the body wall of different type and parenchymal muscles generate bending, shortening, and lengthening movements (Hooge 2001; Tekle et al. 2005; Semmler et al. 2008; Achatz et al. 2010).

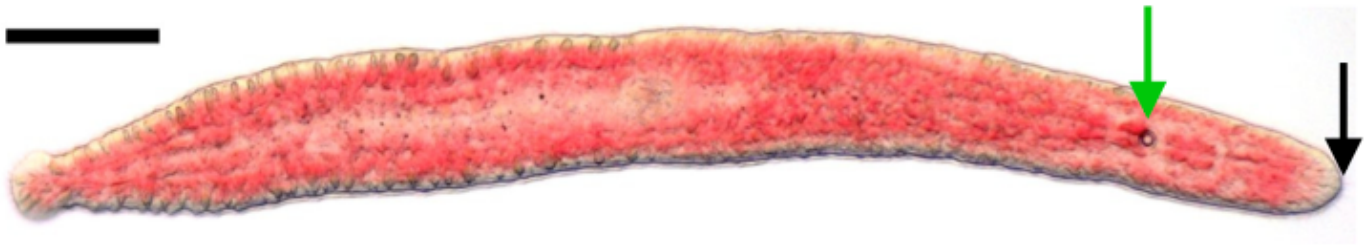


Figure 1.4. The photography of the pink pigmented *Paratomella rubra* from Filey coast in north Yorkshire. (Picture Egger and Tomiczek). The anterior end of a 700µm long adult with the statocyst indicated by a green arrow and a frontal organ indicated by the black arrow. The scale bar is 100 μm.

Several types of sense mechanisms in acoels' bodies direct their movement. Main gravity organ is the statocyst, which is built of the lithocyte including one statolith (see Figure 1.5d). Additionally, acoels have single-celled monociliated receptors, and in some species there are the photoreceptive eyes at the anterior end of the body. The eyespots are not built of ciliary and rhabdomeric elements but are constituted of a pigment cell containing a vacuole with refractive inclusions called concrement (see Figure 1.5c). Their nervous system is formed by a set of longitudinal nerve bundles, which are united by a ring commissure, but do not form an actual brain-like structure. Authors hypothesized that, if acoels are the sister group of all other Blateria, such organization can be a precursor of the cephalization of the nerve system in the ancestor of bilaterians (Perea-Atienza et al. 2015). The food is digested through the mouth, but the position of the mouth is variable from subterminal anterior to the terminal posterior. The pharynges, which are used for sucking the food, are present in some acoel families, but their origin is debated, as there is a lot of variation in their morphology (Todt 2009; Jondelius et al. 2011; Achatz et al. 2013). The gut lacks a lumen (inside space of a cavity, there is no epithelium lined gut) in most investigated species and is therefore commonly termed a central syncytium (multinucleated cells that can result from multiple cell fusions of uninuclear cells) (see Figure 1.6). No typical excretory organs (neohrocytes) have been found in acoels. Acoels are simultaneous or slightly protandric hermaphrodites. In addition, at least some acoels exhibit great regenerative capacity after fission or after experimental amputation (Egger et al. 2007).

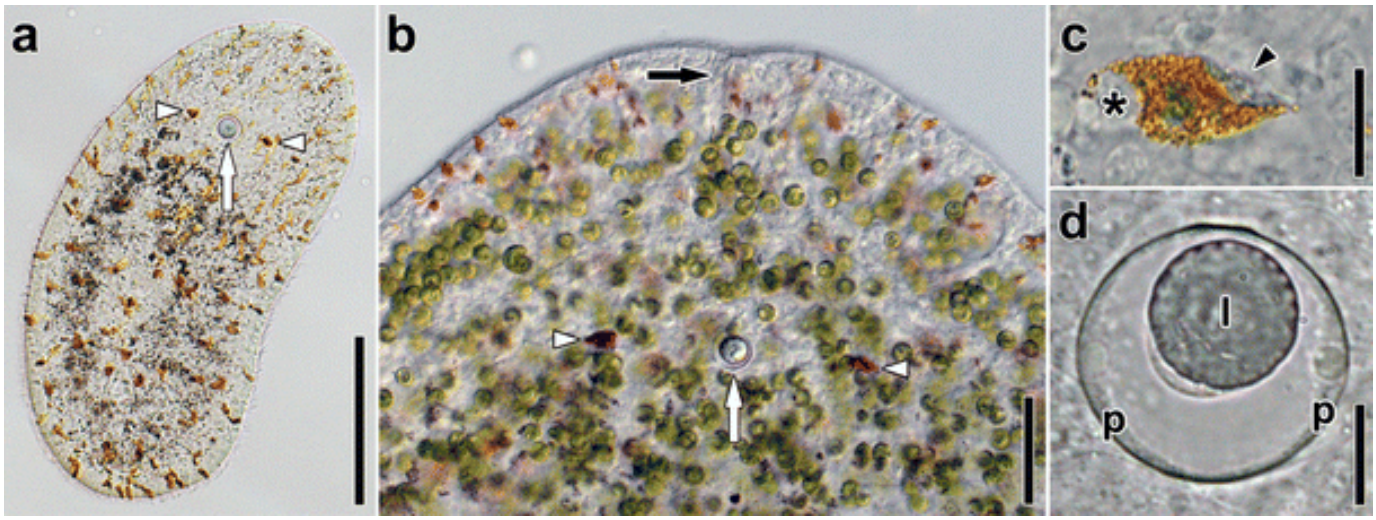


Figure 1.5. Images of sensory structures of live *Symsagittifera roscoffensis* analysed in this study. a) The arrowheads point to eyes, arrow to statocyst. Note absence of symbionts and presence of orange rhabdoids. b) Anterior end of adult with symbionts and rhabdoids. White arrowheads point to eyes, white arrow to statocyst,black arrow to frontal organ. c) Eye of an adult. Asterisk marks nucleus, arrowhead points to eye spots (concrements). d) Statocyst of an adult. Abbreviations: l lithocyte; p parietal cells. Scale bars: a 100 µm; b 50 µm; c 10 µm; d 10 µm (Achatz et al. 2013).

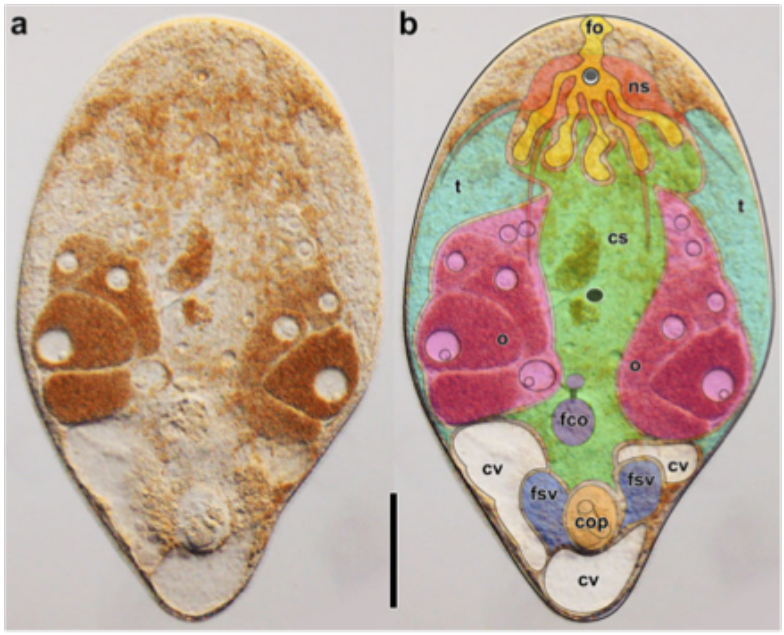


Figure 1.6. Image of a mature and live specimen of *Isodiametra pulchra* analysed in this study without (left) and with superimposed colors (right) to illustrate the general morphology of acoels. From top to bottom: yellow: frontal organ (fo); red: nervous system (ns); green: central syncytium (cs); cyan: testes (t); pink: ovaries (o); gray: mouth; purple: female copulatory organs (fco) composed of seminal bursa, bursal nozzle, and vestibulum (from posterior to anterior); white: chordoid vacuoles (cv); blue: false seminal vesicles and prostatoid glands (fsv); orange: male copulatory organ (cop) composed of muscular seminal vesicle and invaginated penis. Scale bar: 100 µm (Achatz et al. 2013).

1.1.2.2.1 Phylogenetic relationships within the Acoela

Studies based on light microscopy of the copulatory organ lead to the construction of the first family level classification of acoels (Dörjes 1968). Further studies of sperm ultrastructure (Hendelberg 1977; Raikova et al. 2001) and body-wall musculature (Hooge 2001; Tekle et al. 2005) with the use of fluorescent and immunocytochemistry confocal microscopy provided the first hypothesis on the interrelations of these families. This classification was expanded by the analysis based on sequenced studies (Hooge et al. 2002; Jondelius et al. 2011), and lead to a hypothesis that describes the Diopisthoporidae at the base of Acoela, followed by the Paratomellidae and the Prosopharyngida, which are sister group to the group of "higher acoels" (Crucimusculata). *Paratomella rubra* found and sequenced by the joint effort of our lab and my is particularly important for the phylogenetic and gene content analysis as it provides more information about the ancestor of all the acoels not only other sequenced acoels until now (Crucimusculata "higher acoels"). The hypothesis for the evolution of key morphological characteristics, together with the Xenaoelomorph genomes analysed in this thesis indicated on the cladogram (see Figure 1.7).

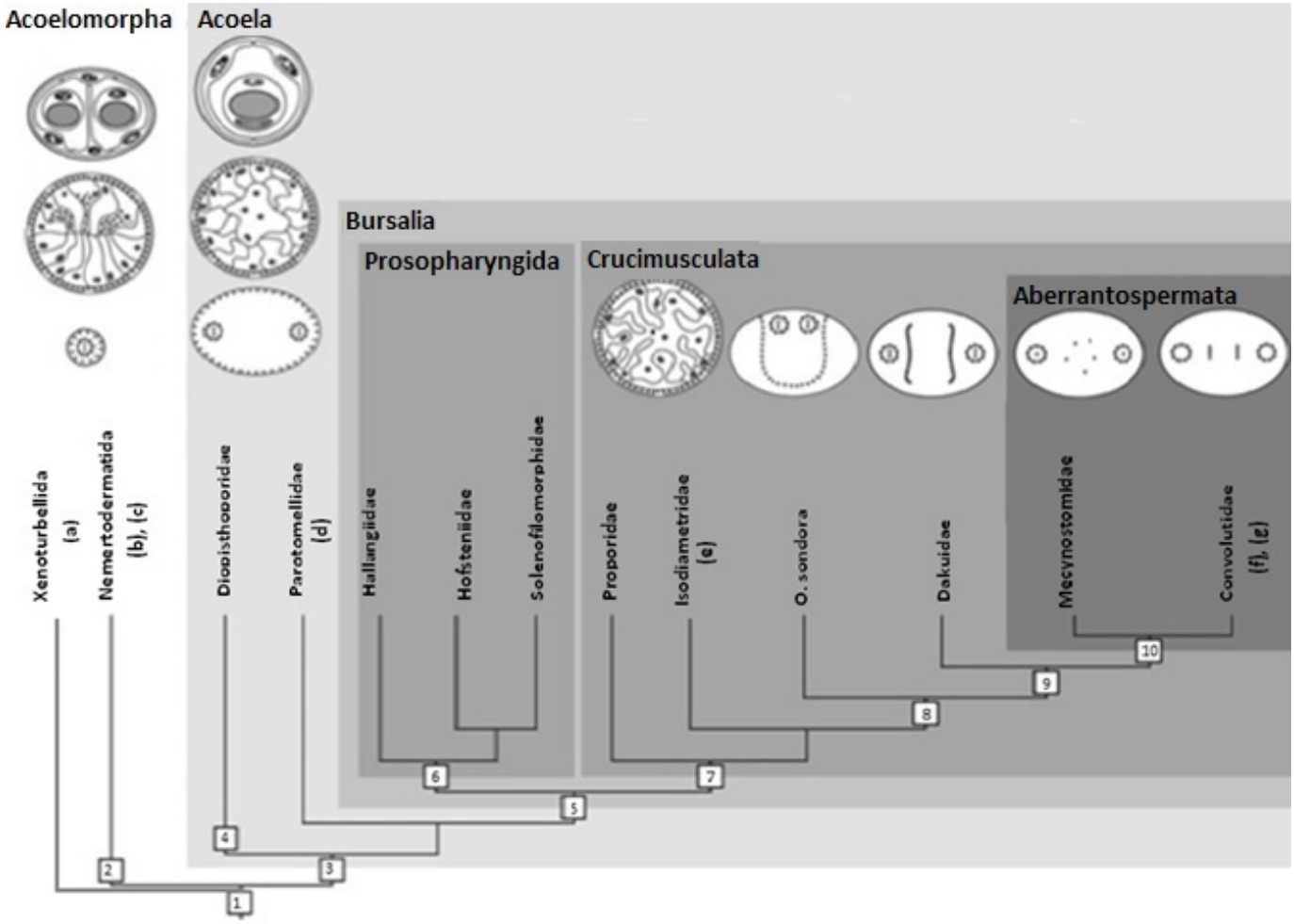


Figure 1.7. Cladogram of the Acoelomorpha with partial family-level systematics of the Acoela. 1. Multiciliated epidermis, ciliary rootlet system, frontal organ, basiepidermal nervous system with ring-shaped brain. 2. Statocyst with two lithocytes (statoliths) and many parietal cells, sperm with cork screw-like morphology. 3. Statocyst with one lithocyte (statolith) and two parietal cells, brain sunk below body wall, lateral fibers at knee of rostral rootlet, biflagellated sperm; digestive system becomes depolarized. 4. Position of mouth at the posterior end. 5. Specialized parenchymal tissue for reception, storage, and digestion of sperm (seminal bursa). 6. Subterminal pharynx at anterior end. 7. Ventral crossover muscles and highly branched wrapping cells. 8. Cytoplasmic microtubules of sperm partially lose contact with membrane and change position toward the center of the cell. 9. Cytoplasmic microtubules of sperm change position toward the center of the cell, stacked bursal nozzles with matrix and gland cells. 10. Central microtubules in axonemes of sperm reduced to allow movement in more than one plane. General scheme after Achatz et al. (2010); schemes of cross sections through statocysts from Ehlers (1985), through bodies after Rieger and Ladurner (2003); systematics and branching after Jondelius et al. (2011). (Fig. Achatz et al. 2013, modified). Species analysed in the study indicated in the cladogram with the letters: (a) *Xenoturbella bocki*, (b) *Meara stichopi*, (c) *Nemertoderma westbladi*, (d) *Paratomella rubra*, (e) *Pseudophanostoma variabilis*, (f) *Symsagittifera roscoffensis*, (g) *Praesagittifera naikaiensis*

1.1.2.3 Nemertodermatida

The acoel's closest relatives are nemertodermatids, the genomes and transcriptomes of two representatives of this clade are analysed in this study (*Meara stichopi* and *Nemertoderma westbladi*). Nemertodermatids live predominantly in marine habitats, mud or muddy sand, and some like *Meara stichopi* lives in a gut of holothurians (e.g. Sea cucumber) (see Figure 1.8). *Meara stichopi* was first found by Sixten Bock and described by Westblad (1949), where *Nemertoderma westbladi* was found by Westblad and described by Steinbock (1938). Nemertodermatids are usually droplet shaped and have a monolayered multiciliated epithelium (tissue that lines the cavities). The cilia comprise from pairs of microtubules and have a characteristic complex rootlet system and tips. There is no continues extracellular matrix (a collection of extracellular molecules secreted by cells that provides structural and biochemical support to the surrounding cells). In most species the mouth opens into a short ciliated pharynx, the gut cavity is not ciliated. There are no protonephridia described. The nerve system is basi-epithelial, with the concentration at the anterior end. The statocyst has two statholiths (Westblad 1937). There is a frontal organ, which consists of gland cells and ciliated cells (Ehlers et al. 1992).

Figure 1.8. Gravid adult of *Meara stichopi* collected in throughout the winter between 2009/2010 - 2013/2014 at collection sites around Bergen, Norway by Aina Børve . The characteristic double statocyst (dst) at the anterior end is indicated (Børve and Hejnol 2014).

1.1.3 Phylogenetic position of Xenacoelomorpha inferred from their morphology

Both Xenoturbellida and Acoelomorpha have been originally placed as the earliest branching Platyhelminthes (flatworms) (Westblad 1949), mainly based on the simplicity of their body plan. Later the similarities in the epidermis of *Xenoturbella* and hemichordates, as well as the similarities in the statocyst found in sea cucumber (Echinodermata) and *Xenoturbella's* statocyst have been noted (Reisinger 1960). The idea of the relation between *Xenoturbella* and Ambulacraria (hemichordates and Echinodermata) have been proposed, as it was hypothesized that *Xenoturbella* might represent the sexually mature larva of an animal related to Ambulacraria. Also later, several authors have referred to *Xenoturbella* as a basal Deuterostome (Bourlat et al. 2003, 2006; Ferrier et al. 2007; Fritzsch et al. 2007; Gee et al. 2003; Perseke et al. 2007). The position of Acoela within the Platyhelminthes also have been questioned (Smith et al. 1986), as it was based mainly on the combination of weak characters, which may be a result of similar habitat: an acoelomate body structure, a densely multiciliated monolayered epidermis, a frontal organ, neoblasts, hermaphroditic reproduction with similar reproductive-organ morphology, biflagellate sperms with inverted axonemes (in acoels and rhabditophorans except macrostomorphans), and lack of hindgut and anus. Further reassessment of morphological evidence within Metazoa, and the phylogeny reconstruction based on large set of morphological characters led to the idea of placing acoels at the base of Bilateria (Haszprunar et al. 1996; Zrzavy et al. 1998). The placement of both Xenoturbellida and Acoelomorpha within Metazoa, strictly based on morphology, was difficult because of lack of strong similarities to other clades, while the placement of acoels shortly became pivotal in the understanding of radial-bilateral transition within animal kingdom.

1.1.4 The hypothesis about evolution of Bilateria

In the 19th century two competing theories, which viewed the Urbilaterian animal (the ancestor of all bilaterians) in a different way and propose different scenarios for radial-bilaterian transition, have been proposed. The first, "acoeloid-planuloid hypothesis" posits that the planulae larvae of cnidarians

transitions into an acoel-like bilaterian animal, which the rest of bilaterian phyla evolved by stages of increased size and complexity (Gaff 1904; Hyman 1959). In this view, coelomate segmented bilaterians derived from simple unsegmented acoelomate creature similar to today's acoelomorphs. The second, "archicoelomate hypothesis " posits a swift transition from either a larval or an adult cnidarian to a complex Bilateria ancestor which already bears through-gut, eyes, coelom and, segments, primitive heart and, very likely, some sort of appendages (Haeckel 1874; Jägersten 1955; Kimmel 1996; De Robertis 1997). In this theory, the acoelomorphs are the state derived by simplification form a coelomate ancestor. Acoelomorpha have an important role in the understanding of the evolution of Bilateria, since some researchers view them as the most similar stage to the ancestor of bilaterians and imagined the Urbilaterian (Last common ancestor of Bilateria) as a creature similar to acoelomorphs. Because of that, the correct placement of Xenacoelomorpha in the animal kingdom is pivotal for better understanding of the early bilaterian evolution.

1.1.5 The history of molecular phylogenetic studies of Xenacoelomorpha

1.1.5.1 Acoelas as basal bilaterians

Previous attempts to place Xenacoelomorpha within the animal kingdom based on morphological similarities were unsuccessful, because of the few morphological characteristics to compare. Authors used molecular phylogenetic analysis to help to resolve the uncertain phylogenetic position of Acoelomorpha and Xenoturbellida. Even though both groups were originally considered to be a part of the Platyhelminthes (flatworms), nucleotide sequence data indicated that neither group is closely related to this phylum (Littlewood et al. 2001; Ruiz-Trillo et al. 2002; Telford et al. 2000, 2003; Jondelius et al. 2002; Philippe et al. 2007; Wallberg et al. 2007). The first argument for the placement of the acoels away from Platyhelminthes came with molecular phylogeny of 61 species, based on the analysis of 18S ribosomal DNA genes sequences, found acoels branching as a sister group to all other bilaterians

(Ruiz-Trillo et al. 1999). More support for this position came with the comparative analysis of mitochondrial code, Telford and supporters noticed significant differences in the mitochondrial codon AUA codes for Ile rather than the normal Met and the mitochondrial codon AAA codes for Asn rather than Lys in acoels and other flatworms. They suggested that the basal position of the aceols is the one of the most parsimonios explanations for the observed changes in the mitochondrial code.

However, while acoels were placed outside bilaterians, nemertodermatids were still remaining in the Platyhelminthes (Ruiz-Trillo et al. 1999; Peterson and Eernisse 2001). Consequently, the evidence from the maximum parsimony and maximum likelihood analysis of 18S rDNA and mitochondrial genes, which included sequences from three species of nemertodermatids, placed both acoels and Nemertodermatida outside of other Bilateria (Jondelius et al. 2002) (see Figure 1.9a), away from Platyhelminthes as separate clades, with Nemertodermatids branching closer to other Bilaterians (Nephrozoa) (Jondelius et al. 2002). However, because acoels and nemertodermatids share many morphological characters (ciliary rootlet system, cilia with shelf-like termination) such a placement was not commonly accepted. Further analysis of large and small subunit ribosomal RNA found no statistical support in favor of paraphyletic relationship between Acoela and Nemertodermatida, and concluded monophyletic Acoelomorpha (Telford et al. 2003). Acoels bear the primitively minimal set of Hox genes (group of genes that control the body plan during embryonic development), which includes only one Hox gene of each class (anterior, central, and posterior) (Cook et al. 2004; Fritzsch et al 2007; Ferrier et al 2007). This was interpreted as a basal characteristic within Bilateria based on parsimony, because other bilaterians possess more Hox genes it was concluded that they were gained after Acoelomorpha-Nephrozoa divergence (Acoelomorpha and other bilaterians). Additionally, acoels lack key microRNAs necessary for organogenesis such as miR-1 (heart) or miR-9 (brain), which was interpreted as a ancestral state within Bilateria (Sempere et al. 2006, 2007).

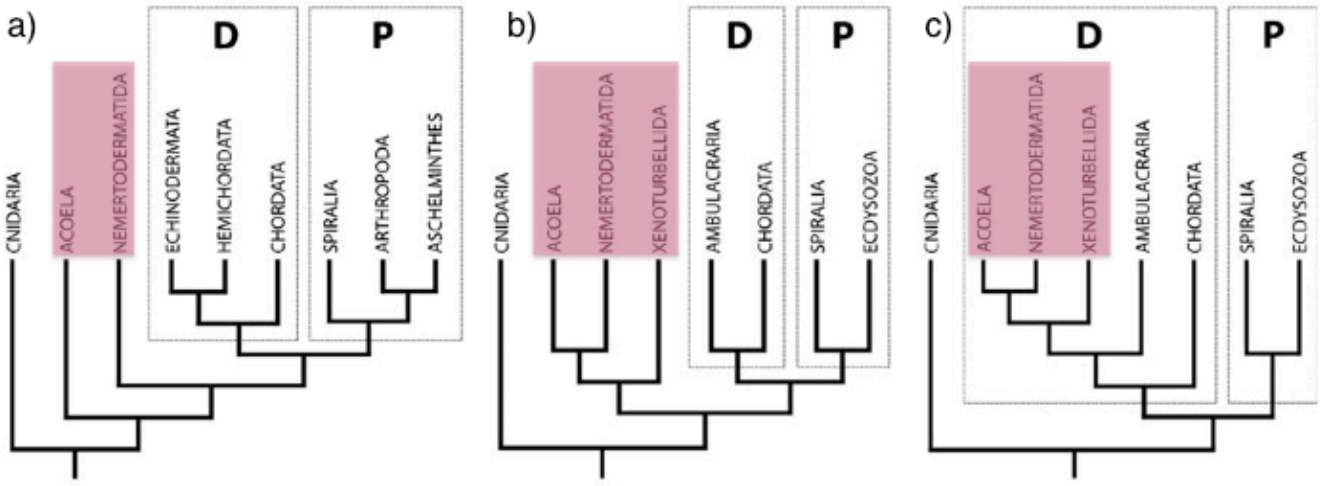


Figure 1.9. Three scenarios for the phylogenetic position of Acoela, Nemertodermatida and Xenoturbellida. a) Both Acoela and Nemertodermatida branching at the base of Bilateria, with Nemertodermatids branching closer to other Bilaterians (Nephrozoa– Protostomia (P) and Deuterostomia (D)) (Jondelius et al. 2002; Wallberg et al. 2007) (with Xenoturbella not included in the analysis but considered as part of Platyhelminthes within Spiralia) b) clade Xenacoelomorpha (Acoela, Nemertodermatida and Xenoturbellida) sister group to other Bilateria (Nephrozoa – Protostomia (P) and Deuterostomia (D)) (Hejnol et al 2009; Srivastava et al. 2014; Cannon et. al 2016) c) clade Xenacoelomorpha (Acoela, Nemertodermatida and Xenoturbellida) sister group to Ambulacraria (Nakano et al. 2013; Bourlat et al. 2006; Telford et al. 2008; Philippe et al. 2009, 2011).

1.1.5.2 Xenoturbella eats molluscs

The first molecular studies of *Xenoturbella* placed the species within the molluscs, based on the analysis of 18S ribosomal RNA and COI genes (Israelsson 1997; Norén and Jondelius 1997). Nevertheless, the lack of similarities to adult molluscs was striking, and was followed by the discovery of the contamination in the *Xenoturbella*'s genetic material with molluscan DNA (Bourlat et al. 2003). Bourlat et al. made efforts to amplify SSU and Cox1/Cox2 genes through PCR amplification from nucleic acids derived from tissue excluding the gut, since *Xenoturbella* eat bivalve molluscs. The phylogenetic analysis of this genes showed that *Xenoturbella* is a deuterostome, related to the Ambulacraria (echinoderms and hemichordates), which explains previously noticed similarities to hemichordates (Reisinger 1960). Bourlat et al. also found that the sequence and orientation of the four genes (Cox1 S2 D Cox2) was only found in the deuterostomes (Bourlat et al. 2003), but also noticed that the typical for Ambulacraria mitochondrial genetic code (AUA codes for isoleucine rather than the methionine) is not present in *Xenoturbella*. Additionally, immunocytochemical data supports the position of *Xenoturbella* close to Ambulacraria, as the antibody found by Stach (Stach et al. 2005) specifically recognizes SALMF Amide-2 short neuropeptide in echinoderms, hemichordate and *Xenoturbella*, but not in other metazoans. The improved phylogenetic analysis of not two but 170 nuclear and 13 mitochondrial genes reassigned *Xenoturbella* as a sister group of Ambulacraria (hemichordates and echinoderms) within the deuterostomes (Chordata and Ambulacraria) (Bourlat et al. 2006). The result of these studies was later confirmed by extended phylogenetic analyses (Philippe and Telford 2006; Telford 2007; Dunn et al. 2008). However, further genome structure investigation of the gene order in the mitochondrial genome (Bourlat et al. 2009), as well as, a phylogenetic analysis of mitochondrial genes (Perseke et al. 2007; Bourlat et al. 2009) indicated a basal deuterostome placement of *Xenoturbella.*

1.1.5.3 Xenacoelomorpha are sister group to Ambulacraria

Based on the previously noticed similarities between acoels and *Xenoturbella*, and previous results regarding *Xenoturbella* (Bourlat et al. 2003, 2006) a new position of Acoelomorpha and Xenoturbellida as a sister group of Ambulacraria has been hypothesized (Telford et al. 2008). Telford noticed that Xenacelomorpha are relatively a long branch, due to the high substitution rate, and the placement at the base of Bilateria could be a result of a Long Branch Attraction artifact (systematic error whereby distantly related lineages are incorrectly inferred to be closely related and are attracted to the base of a phylogenetic tree because both have undergone a large amount of change) (Telford et al. 2008). The recent computational analysis of 197 genes from the EST data and mitochondrial genes, that uses CAT+GTR+Γ model, which categorizes the sites in the sequence alignment based on the amino acid frequencies and substitution rate. This model fits the data better and minimizes the effect of the Long Branch Attraction artifact (the artifact which causes the artificial grouping of fast evolving taxons on a phylogenetic tree (Lartillot et al.)). This analysis places not only Xenoturbellida but also Acoels and Nemertodermatids together as a subclade of Deuterostome called Xenacoelomorpha, a sister group to Ambulacraria (Philippe et al. 2011) (see Figure 1.9c). Moreover, Philippe et al. analysed miRNA content in Bilateria genomes and found that there is a specific miRNA (miR-103) that is specific only for deuterostomes, and can be found in both acoels and Xenoturbella. Furthermore, there is a single miR-2012 that is specific for Ambulacraria and Xenacoelomorpha, and two (XANov-1, XANov-2) miRNAs that are present specifically only in Xenacoelomorpha. Other clade specific genetic feature of deuterostomes that can be found in Xenacoelomorpha is gene, coding for the sperm protein RSB66 (Philippe et al. 2011). Xenacoelomorpha lack several HOX genes (Cook et al. 2004; Fritzsch 2008) and miRNAs (Philippe et al. 2011; Sempere 2007) typical for Bilateria. The lack of this molecular markers suggest that the absence of HOX genes and miRNAs can be linked to a simple morphology of Xenacoelomorpha. Even thought the absence of HOX genes and miRNAs was previously interpreted as a support for the position of Xenacoelomorpha as basal bilaterians (Cook et al. 2004; Sempere 2007), with a Xenacoelomorpha being sister to Ambulacraria it is more understandable that

they may have been lost from a Xenambulacraria Last common ancestor (the ancestor of Ambulacraria and Xenacoelomorpha) as a result of the secondary simplification of a body plan.

1.1.6.4 Xenacoelomorpha are basal bilaterians

Contrary to the results by Bourlat and Philippe (Bourlat et al. 2006; Philippe et al. 2011), another phylogenetic analysis of 94 taxa, in which includes new EST data (expression sequence tags) from 2 nemertodermatids and 3 acoels, concludes a placement of Xenacoelomorpha at the base of Bilateria (Hejnol et al. 2009) (see Figure 1.9b). The tree was inferred based on 1487 orthoMCL gene clusters using Maximum Likelihood method in RAxML, and shows high bootstrap supports for the position of Acoelomorpha at the base of Bilateria, as well as moderate support for the position of Xenoturbellida as a sister group to Acoelomorpha. Hejnol recognizes Acoelomorpha as a relevant outgroup to Nephrozoa (proteostomes and deuterostomes), and based on similarities with cnidarians and ctenophore, considers ancestor of all Bilateria as a simple creature with single body opening and orthogonal nerve organization (consisting of multiple longitudinal dorsal and ventral cords). The week point of the analysis presented by Hejnol was that the position he inferred was only calculated with ML approach, and following analysis by Philippe had less missing data and used better fitted CAT+GTR+ Γ model. However the following analysis, which provides the phylogenetic analysis with 212 gene clusters, and includes 4 nemertodermatids, 7 Acoelomorpha and *Xenotrbella bocki* with both Baysian CAT+GTR+ Γ model and Maximum Liklihood analysis concludes the basal position of Xenacoelomorpha within Bilateria (Cannon el al. 2016). Additionaly, phylogenetic analysis of new *Xenanoturbella* species, together with *Xenoturbella bocki*, a single acoel *Hofstenia* and 16 other metazoan also supports the basal position of Xenacoelomorpha within Bilateria, but only with the Maximum Likelihood approach (Rouse et al. 2016). Within the same paper, the Bayesian analysis of 1,178 genes supports the basal proteosome position of Xenacoelomorpha, while the analysis of mitochondrial proteins again supports the basal Deuterostome position of Xenacoelomorpha.

1.1.6.5 Two conflicting ideas

Previous phylogenetic analysis of the orthologous genes inferred from the ESTs and genomic data using orthoMCL clustering approach and curated reciprocal approaches are not with agreement with each other, likely because authors use different models of a amino acid substitutions and different methods for phylogeny inference (Maximum Liklihood and Bayesian) (Philippe et al. 2011; Hejnol et al. 2009). The recent analysis of mitochondrial proteins, which includes new *Xenoturbella profunda* species (Rouse et al. 2016), supports the basal Deuterostome position and is in agreement with the mitochondrial protein analysis by Bourlat (Bourlat et al. 2006). Other analysis performed by Rouse with site heterogeneous model of evolution (CAT+GTR+ Γ) of 1,178 genes nuclear genes support the position of Xenacoelomorpha as a sister group to protostomes (Rouse et al. 2016). Only the Maximum Liklihood inference with GTR+ Γ model supports the basal Bialateria position of Xenacoelomorpha and is in agreement with the other large scale Maximum Liklihood with GTR+ Γ analysis by Hejnol (Hejnol et al. 2009). However, recent findings by Hejnol group show that they can infer basal bilaterial placement of Xenacoelomorpha with both Maximum Likelihood and Bayesian under CAT+GTR+ Γ; GTR+ Γ and LG+I+ Γ models (Cannon et al. 2016). They obtained their results using small number of 212 genes and used HMM (Hidden Markov Model) gene profile approach. Taking into account that the orthology inference method can influence the outcome of the phylogenetic analysis by including paralogs. Small gene dataset or poor quality alignment can enhance systematic errors or artifacts such as Long Branch Attraction. We aim to improve our previous analysis by using more genes (see Chapter 6) and selecting the best method of finding orthologs from animal genomes (see Chapter 4).

More analysis is necessary to resolve the phylogenetic position of Xenacoelomorpha. Their phylogenetic position is crucial to understand the evolution of early Bilateria, as well as the reasons for Xenacoelomorpha morphological simplicity. Depending on the placement of Xenacoelomorpha on a tree of life, we can hypothesize both on the appearance of the ancestor of all Bilaterians and the way Xenacoelomorpha simplified or kept relatively body plan through evolution. The two possible

locations for the Xenacoelomorpha within the animal tree should be considered i) as the sister clade to all other bilaterians (Wallberg et al. 2007; Jondelius et al. 2002; Littlewood et al. 2001; Ruiz-Trillo et al. 2002; Telford et al. 2000; Telford et al. 2003; Hejnol et al. 2009; Srivastava et al. 2014; Cannon et. al 2016) and ii) as deuterostomes, most closely related to the Ambulacraria (echinoderms and hemichordates) (Nakano et al. 2013; Bourlat et al. 2006; Telford et al. 2008; Philippe et al. 2009, 2011). Below, I will describe the two different views on the evolution of Xenacoelomorpha depending on their phylogenetic position, and two different hypotheses on the appearance of the ancestor of all Bilateria.

1.1.7 Evolutionary implications of phylogenetic positions of Xenacoelomorpha at the base of all Bilateria.

Based on the phylogenetic position of Xenacoelomorpha at the base of Bilateria, supported by previous molecular evidence (Jondelius et al. 2002; Littlewood et al. 2001; Ruiz-Trillo et al. 2002; Telford et al. 2000, 2003), some authors consider Xenacoelomorpha as representative state of early bilaterians, and are regarded as morphologically closest relatives to the Bilateria last common ancestor (Haszprunar et al.; Bagun J, Riutort M (2004)). However, this means that Xenacoelomorpha evolved through a very long branch leading from last common ancestor of all Bilateria, Urbilateria, indicating many sequence changes from Urbilateria state. The characteristics typical for acoels, like no coelom, intra-epidermal nervous system and a simple body plan, can be primitive characteristics for Bilateria last common ancestor. This view supports the planuloid-acoeloid theory, where the Bilateria last common ancestor is regarded as a simple non-segmented creature, similar to the present day acoelomorph worms (Nielsen 2008), but already triploblastic (a mesoderm as well as ectoderm and endoderm) and bilateraly symmetrical (left/right symmetry).

On the other hand Deuterostome-Protostome last common ancestor is supposed to be a very complex animal considering the characteristics that all Deuterostome and Proteostome have, and have through gut, a circulatory system with pumping organ, coeloms and segmentation

(Xavier-Neto et al. 2007). It is not clear however, if Urbilaterian ancestor of Xenacoelomorpha and Deuterostome-Protostome last common ancestor was also fairly complex. If it was, Xenacoelomorpha must have lost these common characteristics on the long branch leading from Urbilaterian ancestor to Xenacoelomorpha. Alternatively, characteristics typical for acoels (intra-epidermal non centralized nervous system, no thought gut and a simple body plan) can be primitive characteristics for Urbilaterian last common ancestor (Hejnol and Martindale 2008), which would involve significant development from Urbilaterian last common ancestor to Deuterostome-Proteostome last common ancestor. In this thesis I am interested, if one of the two scenarios can be supported with the observed significant gene gain from inferred genome content of Urbilaterian last common ancestor to Deuterostome-Proteostome last common ancestor. Gene loss from Urbilaterian last common ancestor to Xenacoelomorpha would be however more difficult to detect, because the inferred gene content of Urbilaterian last common ancestor would be reconstructed based on gene content of Xenacoelomorpha and Deuterostome-Proteostome last common ancestor. The complexity of Urbialteria is debated in the literature, but understanding the gene content of ancestral Bilateria and Deuterostome-Proteostome (Nefrozoa if Xenacoelomorpha are basal Bilateria) if important for understanding the early evolution of bilateral animals.

1.1.8 Evolutionary implications of phylogenetic positions of Xenacoelomorpha as a sister group to Ambulacraria.

The position of Xenacoelomorpha as a sister group of Ambulacraia implies that they evolved from the same common ancestor as chordates, hemichordates and echinoderms (deuterostome last common ancestor). Typical characteristics for deuterostomes are gill slits, endostyle and postanal tail, as well as typical features of embryonic development such as radial cleavage and deuterostomy, together with other bilaterian characteristic such as through gut, eyes and cephalic brain (Gerhart et al. 2005). There are no signs of gill slits, endostyle and anus in adult Xenacoelomorpha, and the cleavage is not radial, as well as no typical bilaterians characteristics are present in Xenacoelomorpha. Therefore, these characteristics

must have been lost from deuterostome last common ancestor as a result of secondary simplification. I am interested, if there is any signs of genetic characteristics that can be correlated with phylogenetic position of Xenacoelomorpha within deuterostomes, or more precisely as a sister group of Ambulacraria. These genetic characteristics would be, genes specific to deuterostomes or Ambulacraria present in Xenacoelomorpha, simultaneous gene losses in Xenacoelomorpha and Ambulacraria, simultaneous gene losses in Xenacoelomorpha and other deuterostomes.

1.1.9 Outline

To get more evidence on the controversial phylogenetic position of Xenacoelomorpha within Metazoa and to understand the genetic correlates of the morphological simplification of Xenacoelomorpha, I first gathered the new sequence data from genomic and transcriptomic assemblies of the seven Xenacoelomorpha species thanks to Xenocoelomorpha Genome Project 2014 (*Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Praesagittifera naikaiensis, Pseudophanostoma variabilis, Paratomella rubra* and *Xenoturbella bocki*) and 60 other animals from online resources. I used these resources for family content and phylogeny inference. In Chapter two, I characterize the quality and the completeness of these assemblies and use it to construct 7 new Xenacoelomorpha proteomes (entire sets of proteins expressed by a specific organism (UniProt Consortium, 2010)), which are the subject of further analysis. Next, in Chapter three, I used PhylomeDB database to create the sets of clade specific and ancestral gene families. I developed new algorithm termed family-RBH, which allowed me to investigate the presence of these families in 7 Xenacoelomorpha proteomes. In Chapter four, to choose the best method for constructing phylogenetic matrix, I compared three commonly used methods for inferring groups of orthologous genes, using lophotrochozoan genomic data as a test dataset. I developed a phylogenetic pipeline for the species phylogeny reconstruction from the orthology groups. I tested the performance and the applicability of these three methods for the reconstruction of the molecular phylogeny from genomic

data and choose the best method for the inference of Xenocoelomorpha position on the animal tree of life. In Chapter five I used the proteomic data from 67 species (58 Metazoan including 7 Xenacoelomorpha) to create the database of gene families using OMA standalone. I analysed the content of these families by investigating simultaneous gene losses in Xenacoelomorpha and other clades. I reconstructed gene evolutionary events (loss, duplication, de novo creation) within these families, which allows me to quantitatively follow gene evolution across Metazoa. In Chapter six I used the phylogenetic pipeline established before to perform large-scale molecular phylogenetic analysis of Metazoa and investigate the phylogenetic position of Xenacoelomorpha. The obtained results were summarized and support the position of Xenacoelomorpha as a sister group of Ambulacraria. However, my results question the monophyly of Deuterostomes, and support the hypothesis that Bilateria diverged in short period of time into three monophyletic linages (Chordata, Xenambulacraria and Protostomia), which could explain previously obtained, contradictory results by other authors (Bourlat et al. 2006; Philippe et al. 2011; Hejnol et al. 2009; Cannon et al. 2016; Rouse et al. 2016). In that case characteristics, such as gill slits, endostyle or deuterostomy also not present some protostomes and Xenacoelomorpha would be basal bilaterians characteristics. Both Protostomia and Xenacoelomorpha lineages could have lost them during the course of evolution independently.

Chapter 2

Quality assessment of the Xenacoelomorpha genomic and transcriptomic sequences

2.1 Introduction

The evolution of Xenacoelomorpha is a subject of recent debate in biology, which affects the understanding of the origins of Bilateria (Hejnol et al. 2009; Srivastava et al. 2014; Philippe et al. 2011; Telford 2013; Cannon et al. 2016; Rouse et al. 2016). Here, we aim to better characterize the relation of Xenacoelomorpha with other clades, by analyzing the new genomic and transcriptomic sequences of the acoels *Symsagittifera roscoffensis, Pseudophanostoma variabilis, Paratomella rubra* and *Praesagittifera naikaiensis* (no genomic sequencing was performed for *Praesagittifera naikaiensis*); the nemertodermatids *Meara stichopi* and *Nemertoderma westbladi* and the xenoturbellid *Xenoturbella bocki*, using this new data we construct 7 new proteomes (entire sets of proteins expressed by a specific organism (UniProt Consortium, 2010)) (see Figure 2.1).

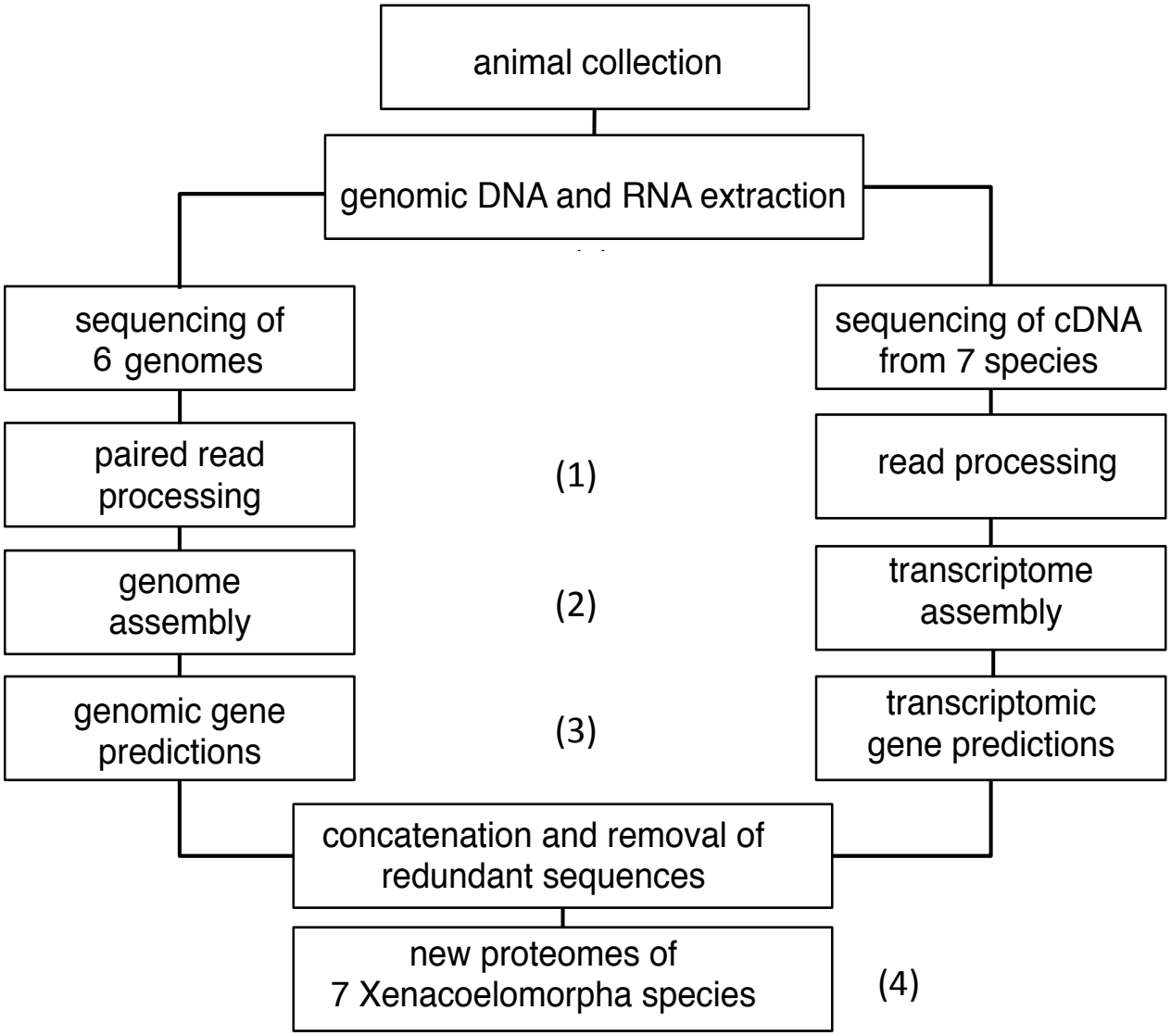


Figure 2.1. The block scheme represents the workflow of sequence data processing for 7 sequenced species from Xenacoelomorpha clade (no genomic sequencing was performed for *Praesagittifera naikaiensis*). The quality tests were applied on read level (1), assembly level (2), genomic and transcriptomic gene predictions level (3) and complete protein set for each species (4). From the raw reads (1) I estimated the genome size for each species and estimated the repeat content and heterozygosity for each pair of sequencing reads. In (2) I assessed the genomic assembly quality by calculating statistics of the genomes contiguity. (3)I check the number of genomic and transcriptomic gene predictions for each species and analyse the content of core animal proteins in each of the created proteomes.

Here, we first collected specimens from 7 different species of Xenacoelomorphs and extracted their genomic DNA and mRNA. We gathered 10 genomic libraries of combined 1,317,490,800 of 100nt illumina genomic paired end reads for acoels *Symsagittifera roscoffensis,* 3 libraries of 420,359,700

reads for *Pseudophanostoma variabilis,* 583,378,230 reads for *Paratomella rubra,* 3 libraries of combined 563,402,050 reads paired end reads for nemertodermatids *Meara stichopi* and 3 libraries of combined 724,847,170 reads for *Nemertoderma westbladi* and 6 libraries of 493,800,804 reads for xenoturbellid *Xenoturbella bocki,* (in cooperation with Albert Poustka)(Table 2.2). From the raw reads we managed to assemble 284,097 contigs for *Symsagittifera roscoffensis,* 455,660 contigs for *Pseudophanostoma variabilis,* 1,044,300 for *Paratomella rubra,* 5,500,396 contigs for *Meara stichopi,* 1,474,389 contigs for *Nemertoderma westbladi* and 106,304 contigs for *Xenoturbella bocki* using SOAPdenovo. From that, we predicted 113,993 ORFs (Open Reading Frames) for *Symsagittifera roscoffensis,* 115,245 ORFs for *Pseudophanostoma variabilis*, 52,346 ORFs for *Paratomella rubra,* 130,115 ORFs for *Meara stichopi,* 80,966 ORFs for *Nemertoderma westbladi,* 21,769 ORFs for *Xenoturbella bocki.* Additionally, we assembled the transcriptomes for each species using Trinity (chosen based on previous experience with *Maritigrella crozieri* (Lapraz et al. 2013)) and predicted 18,495 ORFs for *Symsagittifera roscoffensis,* 32,043 ORFs for *Meara stichopi,* 18,968 ORFs for *Nemertoderma westbladi*, 23,209 ORFs for *Xenoturbella bocki*, 22,287 ORFs for *Pseudophanostoma variabilis*, 25,703 ORFs for *Praesagittifera naikaiensis,* and 28,881 ORFs total for *Paratomella rubra.* We observed that on average 24% of genes might undergo alternative splicing, with an average 1.3 isoforms per unigene (Table 2.2). We translated both gene predictions into protein sequence and joined both predictions and clustered using CD-HIT with a 97% identity threshold (Fu et al. 2012), which resulted in non-redundant proteomes for each species. We obtained 32,456 complete gene predictions in *Symsagittifera roscoffensis,* 35,867 complete gene predictions in *Meara stichopi,* 23,233 complete gene predictions in *Nemertoderma westbladi,* 27,378 complete gene predictions in *Pseudophanostoma variabilis,* 24,329 complete gene predictions in *Paratomella rubra,* 19,206 complete gene predictions in *Xenoturbella bocki* (see Figure 2.1).

In the consecutive Chapters (3-6), we will make use of this new proteomes of Xenacoelomorpha presented here (ideally, the protein sequences associated with every protein-coding gene in all genomes)

to analyse the gene family content (see Chapter 2,4) and construct amino acid supermatrices for the inference of Xenacoelomorpha phylogenetic position on the animal tree of life (see Chapter 5). For this purpose the quality of the sequence data is essential to understand the outcome of future analysis. Here, I describe the process of generating the Xenacoelomorpha protein sets from genomic and transcriptomic resources and characterizing their quality of this data on multiple levels (read level (see Figure 2.1 (marked with (1)), assembly level (marked with (2), gene prediction level (marked with (3)). First, I will describe the potential problems with the sequencing data that could appear on each of the levels (see Section 2.1.1 (levels marked with 1,2,3 on Figure 2.1), and describe the quality tests I applied on each level (see Section 2.2.1).

2.1.1 Potential factors influencing the quality of high-throughput sequencing data

Recently, Next Generation Sequencing (NGS) technologies have improved the accessibility of new genomic data from non-model organisms with relatively low cost and remarkable speed. However, non-model organisms often lack a reference genome, making quality control very challenging, in comparison to cases where reference genomes are available (Trivedi et al. 2014). I will characterize the potential factors that could influence the quality of the Xenacoelomorpha genomic and transcriptomic assemblies and in consequence the proteomes, which were constructed based on these assemblies. The errors and the artefacts can be introduced during the process of reconstructing the protein content of extant non-model organisms such as 7 sequenced xenacoelomorphs during DNA and RNA extraction, sequencing process, read processing, assembly process and gene prediction process. These factors can influence the quality of the sequence data, and impact the outcome of future analysis of gene content (see Chapter 3,5) and phylogenetic analysis (see Chapter 6).

First difficulties in the process of animal genome or transcriptome reconstruction could occur already during DNA or RNA extraction. Foreign genetic material could be accidently incorporated to the DNA/RNA extract by the researcher sample collection or sequence preparation. Parasites, commensals

and bacteria live on the surface and in the gut of the animal of our interest, and their DNA is often hard to remove during the probe preparation. The presence of foreign genetic material in the sequence reads, not only makes the assembly process more difficult, but also can result in the presence of foreign genome contigs or transcripts in the assembly, and protein predictions. This can influence the result of the future analysis, if not removed beforehand. Current methods for removing the contamination from the sequence data rely on exact matching of short subsequences of *k*-mers to the database (Davis et al. 2013; Merchant et al. 2014; Ramirez-Gonzalez et al. 2013) or clustering analysis of proportion of GC bases and read coverage followed by taxon annotation of the contigs (Kumar et al. 2013). The k-mer approach is limited to the sequences present in the database, unlike the clustering analysis, however the clear separation of contigs based on the GC content and read coverage works well only for removing microbial contamination. We have tried to remove the contamination on multiple levels of our genome analysis. First, we have cleaned the raw sequence reads by classifying them into groups representing the same or similar species using PhymmBL. Second, we have classified the gene predictions from using Kraken database. And last, we have cleaned orthology groups based on phylogenetic trees for each gene and the BLAST matches to National Center for Biotechnology Information (NCBI) database.

The large size of eukaryote genomes makes them challenging to be well reconstructed from short reads. Potential problems are caused by sequencing errors, which can occur during sequencing process. This often happens at the beginning and the end of the reads. Moreover, secondary structures as well as GC-rich regions can be difficult to cover. All this factors make sequence data noisy (Mitchelson et al. 2011). Repetitive DNA, often present in high quantities in eukaryotic genomes, is problematic during the process of assembling sequencing reads into larger fragments (contigs). Microsatellites, low complexity DNA, transposons and retrotransposons are present in the genome in the form of repetitive sequences. Frequent repeats result in biased genome coverage and incomplete assembly. Large tandem repeats are

most difficult for the assembly programs to deal with and often result in premature termination of the contigs (Salzberg and Yorke 2005).

Genome assembly is very challenging for highly polymorphic species. For most species sequenced so far, the data was collected either for haploids or for the populations, which have low effective size or are inbred (Vinson et al. 2005; Mewes et al. 1997; Chinwalla et al. 2002). The rate of the polymorphism, depends on the effective population size, and in insects and marine animals is especially high (two magnitudes higher than human (0.5%)(Pushkarev et al. 2009)). For some non-model organism it is difficult to obtain large enough DNA samples from single individual or clones. As a result, the assembly obtained this way has low quality in locations where the heterozygosity occurs. Most of the assembly programs rely on the construction of a de Bruijn graph (the graph constructed from words of length k in sequence reads and overlaps between k-mers (Zerbino et al. 2008; Compeau et al. 2011)). Above mentioned factors such as sequencing error, polymorphism and repeats cause problems in solving the assembly from the graph (the genome assembly is modelled as the solution for the Shortest Common Superstring (finding shortest circular superstring that contains each substring exactly once) the solution for this problem would represent each of these repeats only once in the assembled genome (Medvedev et al. 2007)) and result in incomplete, fragmented or misassembled sequence. Therefore, solving the superstring problem from the de Bruijn graph tends to be difficult for eukaryotic genomes and results in the collection of unique genome fragments (contigs). Different assembly programs tend to produce different quality genome assemblies, depending on the properties of the genome. The quality of the assembly can be assessed by contiguity (a measure of the contigs lengths in the assembly) and completeness of an assembly (percentage of known sequence).

The accuracy and the performance of the gene prediction methods is influenced by the quality of the genome and transcriptome assemblies. The low quality genomic or transcriptomic assembly affects the ability to predict gene sequence. Fragmented genomic assembly or transcriptomic results partial gene

predictions if the contig ends prematurely. Additionally apparent multiple genes could be created, if two contig broke in the middle of gene sequence, an genome was polymorphic and there were multiple misassembled contigs for one the genome fragment or alternative variants of the same transcripts. Missing data from the assembly or lack of transcript can result in missing genes from the proteome. Moreover, accurate prediction of protein sequence from the genome is a difficult task, even if the whole genome sequence is known (Guigo et al. 2006). Gene prediction methods can be divided into three categories, i) single genome *ab initio* predictors (methods that use statistical sequence patterns, such as the coding reading frame, codon usage or splice site), ii) mapping of these known gene sequences onto the genome sequence and iii) the predictions based on the patterns of sequence conservation between genome sequences of evolutionarily closely related organisms. In my case there are no evolutionarily closely related organisms to compare (iii). We mapped available EST data to genome sequence and used this for the training set for AUGUSTUS, however this resulted in very little gene predictions because of the fragmented genomic assemblies (in cooperation with Albert Poustka). We decided to use GeneScan *ab initio* genomic predictors in further analysis as we were interested in as many predictions as possible, as false positive gene predictions were removed from the analysis in later stages in chapter 5 as they were not grouped to the orthology group and are so called singletons.

2.1.2 Testing the quality of Xencoelomorpha Next Generation Sequencing data

To investigate if the potential problems mentioned above (quality of sequencing reads, assemblies, protein predictions, and the presence of the possible contamination) influence the quality of the new Xencoelomorpha sequence data, we have undertaken the following procedures:

i. We have estimated the size of the genomes based on the read coverage of genes believed to be in single copy (Mi et al. 2013).

ii. We have analysed the quality of the genomic sequencing reads using a probabilistic classifier (Simpson 2014) to characterize the presence of repeat content and heterozygosity.

iii. We have analysed the quality of the genome assemblies by calculating the basic metrics of assembly quality.

iv. We have analysed the count and the completeness of gene predictions from the genome and transcriptome.

v. We investigated the presence of the core Eukaryote proteins in the proteomes.

vi. We sought to identify possible contamination in the genomic and transcriptomic protein predictions.

I will describe the procedures we have undertaken on each step of our quality control process presented above, and how we addressed the potential problems with the data at each level of the protein set reconstruction.

2.1.3 The aims of the quality assessment

First, to better understand the properties of the genetic data we aimed to estimate the read coverage of the sequenced genomes and estimate the genome size. To calculate the mean read

coverage of genome, we have mapped filtered reads to single copy genes (RNA polymerase 2, elongation factor 2, 60S ribosomal protein L18A; 50S ribosomal protein L4). Next, based on the mean read coverage and reads number we estimated the genome size. This simple analysis showed that Xenacoelomorpha genomes, with the genome size similar to other animal genomes, have a high coverage and will be potentially difficult to assemble using standard assembly methods due to high level of sequence repeats and heterozygosity.

Next, we aimed at investigating the ratio of the repeat content and the heterozygosity in the Xenacoelomorpha genomes. One possible tool to perform such an analysis is the sga preqc program (Simpson et al. 2014). The program samples k-mers (words of length k in sequence reads) on the subsets of read-ends and construct de Bruijn graph (directed graph representing overlaps between k-mers) that is part of the assembly process. Next, the program analyses the local structures of the graph using a probabilistic classifier, which allocates the connection between k-mer, based on the shape of the connectivity, into sequence variation, sequence errors and repeats. We show that the Xenacoelomorpha genomes, sequenced in this project (joint efforts of Max Telford and Albert Poustka Lab, sequenced and assembled thanks to Xenocoelomorpha Genome Project 2014), are highly complex and contain a high ratio of sequence repeats and sequence variation.

Additionally, we aimed to evaluate the quality of the Xenacoelomorpha genome assemblies (performed by Albert Poustka and Max Telford) and compare them with the quality of the previously published assemblies of other animal genomes. To do this, we used a set of perl modules from the Assemblathon 2 competition (Bradnam et al. 2013), that we implemented in a perl script to evaluate the basic assembly properties (cooperation with Daniel Jeffares). Using our program we measured the basic metrics of assembly quality, such as N50 contig size, N50 scaffold size, the number of contigs greater than 10kb and the percentage of gaps in the scaffolds. We showed that the Xenacoelomorpha genome assemblies have a low contiguity and that the assembly scaffolds contain a high percentage of gaps. The

study revealed that only the assembly of *Xenoturbella bocki* has a quality comparable with other previously published genomes.

We aimed to construct the protein sequences associated with every protein-coding gene in all genomes (proteomes). In order to do that, we predicted the protein sequences from the Xenacoelomorpha genome assemblies using the GeneScan (Burge et al. 1998) and from the transcriptome assemblies using Trinity (Grabherr et al. 2011). Our results show that the number of protein predictions corresponds to the assembly quality, and is smaller if the genome assembly is better. Moreover, the number of the predicted genes from the Xenacoelomorpha transcriptome assemblies corresponds to an average number of genes in animal genomes. In order to maximize the completeness of proteomes in our future analysis, we joined protein predictions from both the genome and the transcriptome, and remove the redundancy by clustering with using 97% similarity in CD-HIT (Limin Fu et al. 2012). The clustering helped to reduce polymorphic multiple copies of the same gene, however partially overlapping genes from the genome and transcriptome or fragmented genes, which appear as multiple genes were not joined together. We do not use any type of local sequence alignment clustering to join multiple copies of the same gene, as this type of clustering often produces artificial sequences and I decided not to use it. In future phylogenetic analysis (see Chapter 6), I choose the longest or the best gene fragment for sequence alignment for the alignment.

Next, we verify the completeness of the constructed Xenacoelomorpha proteomes by identifying the presence of the core Eukaryote proteins. To do that, we used the OMA orthology groups calculated in Chapter 5 (http://omabrowser.org; Roth et al. 2008). Out of these OMA orthology groups, we isolated a subset of 100 core proteins present in at least 51 out of 67 animal species (present in at least 75 %). We measured the proportion of core proteins present in the Xenacoelomorpha proteomes and compared it with other reference proteomes. These analysis reveals that a high proportion of the core proteins are present in the Xenacoelomorpha proteomes and is the same as published for other reference protein sets

(http://www.ncbi.nlm.nih.gov/refseq/). Furthermore, we show that the use of multiple proteomes of Xenacoelomorpha improves the ability to find core proteins in at least one of the proteomes, and allows us to be more confident about the gene content inferred at the Xenacoelomorpha last common ancestor (XLCA).

Here we show that genomes are highly heterozygous and contain large number of repeats, which resulted in low quality assemblies. Protein sets constructed based on the genomic and transcriptomic gene predictions are fragmented, which results in predicting many more presumable gene sequences that an average animal gene count. However, we were able to find high proportion of the core proteins is present in the Xenacoelomorpha proteomes. Furthermore, we show that the use of multiple proteomes of Xenacoelomorpha improves the ability to find the core proteins in one of the proteomes even further, and allows us to have a good confidence in the gene content inferred at the Xenacoelomorpha last common ancestor (XLCA). Additionally we have identified foreign species contamination in the protein datasets we have constructed. We plan to further improve the quality of Xenacoelomorpha proteomes, by removing contaminated sequences from the protein datasets. Additionally, we plan to identify split genes in fragmented genome assemblies, by using ESPRIT software and link unassembled genomic segments together, to improve the genome assemblies.

2.2 Methods

2.2.1 Animal collection, DNA/mRNA extraction and sequencing

We collected the sand containing adult specimens of the acoel *Paratomella rubra* from the upper intertidal zone along the beach in Filey in North Yorkshire. The sand was washed with 5mM $MgCl_2$ solution, in order to anaesthetize the worms without killing them and the wash was filtered through a microporous cellulosic membrane (0.2-*μ*m pore size). First, pink worms were separated from the rest of the fauna under the binocular. The morphology of 400 specimens was confirmed under the light microscope, and the worms were frozen in -80°C. The Genomic DNA was extracted using the QIAamp kit following the manufacturer's protocol. The mRNA was extracted using a standard guanidinium thiocyanate-phenol-chloroform extraction protocol. The Genomic DNA was sequenced using the NGS Illumina technology, producing six 2x100bp paired-end libraries, with insert sizes comprised between 300 and 450 bp (as estimated from an electrophoresis run using High Sensitivity D1K ScreenTape).

The Genomic DNA of *Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Xenoturbella bocki, Pseudophanostoma variabilis,* were sequenced using the NGS Ilumina technology, producing 2x100nt paired-end reads (data provided by Albert Poustka).

2.2.2 Read processing and the genome complexity analysis

The Sequencing reads were trimmed on both ends based on the base quality using Qtrim software (Shrestha et al. 2014). Reads with GC content greater than 55% were discarded, on the assumption that they were bacterial contaminants (Kumar et al. 2013). Paired reads libraries were sampled using the sga preqc software for each paired library separately (Simpson 2014). The software samples k-mers (words of length k in sequence reads) on the subsets of reads and construct de Bruijn graph (directed graph representing overlaps between k-mers). Next, the program explores the local

structures of the graph using a probabilistic classifier. The classifier allows us to distinguish between sequence variation, sequence errors and repeats.

2.2.3 Genome size estimation

The Genome size estimates were obtained by mapping the reads on four putatively single-copy genes (chosen based on literature search and confirmed using PANTHER database (Mi et al. 2007)): elongation factor 2, 60S ribosomal protein L18A; 50S ribosomal protein L4, following an unpublished approach currently being developed by Jens Bast at the University of Göttingen (Jean-François Flot, personal communication). We confirmed the presence of a single copy orthologs of the target genes using the PANTHER family database (Mi et al. 2007). The DNA sequences of these genes were identified by the reciprocal best BLAST hit in the *Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Xenoturbella bocki, Pseudophanostoma variabilis, Paratomella rubra* assemblies. The read library, previously trimmed with Qtrim software (Shrestha et al. 2014) and filtered for PCR duplicates with samtools rmdup (http://samtools.sourceforge.net/) read library reads were mapped on each selected contig using bowtie2 (Langmead and Salzberg 2012). Based on the count of well-mapped reads we calculated the coverage depth for each pair of the Xenacoelomorpha ortholog and read library. The Genome size was calculated according to the formula Genome size = (number of reads * read length / 2 * coverage depth) (Jens Bast, (http://www.jensbast.com/?page_id=82); Jean-François Flot personal communication) for each marker in each species. The means of 4 genome size estimations for each species was presented as a final result.

2.2.4 The Genome Assembly and assembly properties

Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Xenoturbella bocki, Pseudophanostoma variabilis, Paratomella rubra genomes were assembled from shotgun reads, using the SOAPdenovo2 assembler (Luo et al. 2012). A repeat library was created using RepeatScout (Price et

al. 2005), the interspersed repeats and low complexity DNA sequences were masked using RepeatMasker (Tempel et al. 2012; Tarailo-Graovac et al. 2009).

We investigate the distribution of contig lengths in the genomic assemblies using a basic set of metrics from Assembathon2 competition (Bradnam 2013). We have modified assemblathon_stats.pl perl script to evaluate the assembly properties without reference. We distinguish between contigs and scaffolds based on the content of unknown nucleotides in the sequence (containing a stretch of unknown nucleotides (Ns > 5)).

2.2.5 The Transcriptome assembly

Paratomella rubra, Xenoturbella bocki, Meara stichopi, Nemertoderma westbladi, Pseudophanostoma variabilis, Praesagittifera naikaiensis transcriptomes were assembled using the Trinity de novo transcriptome assembly software pipeline (by Max Telford). Open Reading Frames were predicted using the TransDecoder (http://transdecoder.sourceforge.net/).

2.2.6 Assembly decontamination

Assemblies were decontaminated using PhymmBL v 4.0 (Brady et al. 2011) by comparison with reference to the hidden Markov models of DNA (procedure performed by Daryl Domman and Matthew Rowe). The analysis revealed the presence of the contigs from Chlamydiae, Protobacteria and Cyanobacteria in our dataset. The contigs classified by PhymmBL as non-Metazoa were removed from the assembly before the gene prediction process.

Genome and transcriptome nucleotide gene predictions were classified using Kraken to the groups based on the species of origin. For the reference distinct 31-mer library containing genomes from NCBI's RefSeq database was created using Jellyfish multithreaded *k*-mer counter (Marçais at al 2011). The accession number of the gene and Tax ID was stored. The sequences were filtered for non Xenacoelomorpha Tax IDs (Wood et al. 2014).

2.3 Results and discussion

2.3.1 The Xenacoelomorpha genomes are difficult to assemble due to heterozygosity and high repeat content

We have sampled words of length k in sequencing reads (k-mers) using preqc sga software on two sets of paired reads of *Xenoturbella bocki, Meara stichopi, Nemertoderma westbladi, Pseudophanostoma variabilis, Praesagittifera naikaiensis* and reference dataset of genomes provided by Simpson of human, bird, fish, oyster and yeast (Simpson 2014). We constructed the de Bruijn graph representing the overlaps between these words. We have explored the local structures of the assembly graph and classified them into three categories (sequence errors, variation and repeats) (Simpson 2014). Islands of heterozygosity cause a characteristic structure known as “bubbles” in the assembly graph due to allelic differences in the genomes, which are recognized by preqc sga classifier as sequence variation.

The graphs constructed from the *Xenoturbella bocki lane 2 (batch of the reads in the same flow that goes into the sequencing machine), Meara stichopi lane 1,2, Nemertoderma westbladi lane 1,2, Pseudophanostoma variabilis lane 1,2, Praesagittifera naikaiensis* lane 1,2 reads branch more than 1 in 100 vertices due to sequence variation (see Figure 2.1). Reference genomes of human, bird, fish, oyster and yeast as well as lane 1 of the reads from *Xenoturbella bocki* branch less then 1 in 100 vertices due to sequence variation. A frequent branching of the de Bruijn graph due to a sequence variation suggests that there is a high rate of a heterozygous variation in all Xenacoelomorpha genomes. Most of the Xenacoelomorpha assembly graphs (except *Xenoturbella bocki* lane 1) branch due to sequence variation even more frequently than a highly heterozygous oyster genome with 1% heterozygosity rate (Zhang et al. 2012) (and much higher than human with heterozygosity rate 0.1% (Venter et al. 2001), we obtained the same rates as published in the literature when we repeated the analysis for the reference human and oyster genomes (see Figure 2.1)). Single nucleotide polymorphisms and insertions, as well as deletions,

exist in genomes due to the fact that we are not sequencing a single DNA molecule, but a collection of such, from different cells and often multiple individuals (all of the Xenacoelomorpha, except from *Xenoturbella bocki* are very small, and to reach the critical mass of the DNA multiple individuals are required). A high frequency of variant branches in the assembly graph makes the assembly more difficult to resolve, because it is increasing the number of possible walks on the de Bruijn graph that represent the sequence of the genome with many alternatives.

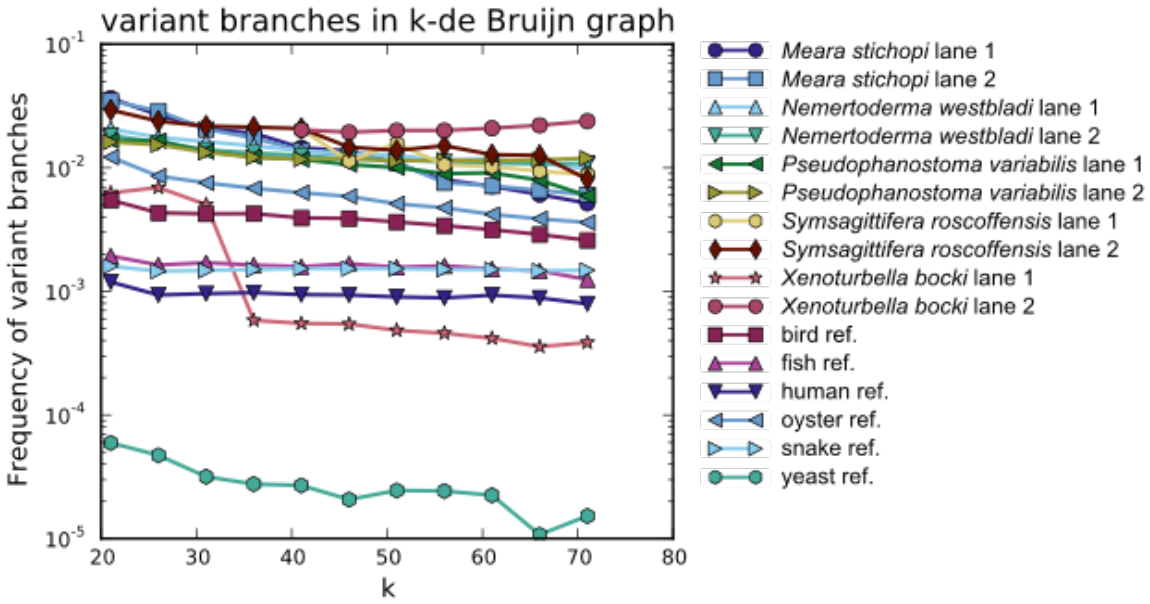


Figure 2.2. High rate of the heterozygous variation in the Xenacoelomorpha genomes in comparison to the reference genome assemblies. Frequency of the de Bruijn graph branching due to the sequence variation as a function of k, measured for two sequencing experiments for each Xenacoelomorpha species. The Xenacoelomorpha genomes (top six samples) show a high frequency of the variant branches compared to the reference sequences from previously sequenced genomes (bottom five).

Another type of branching in the assembly graph is caused by repetitive sequences in the genome. The graphs constructed from the *Xenoturbella bocki lane 2 (batch of the reads in the same flow that goes into the sequencing machine), Meara stichopi lane 1,2, Nemertoderma westbladi lane 1,2, Pseudophanostoma variabilis lane 1,2, Praesagittifera naikaiensis* lane 1,2 reads branch more than 1 in 100 vertices due to sequence repeats (see Figure 2.2). Reference genomes of human, bird, fish, oyster and yeast as well as lane 1 of the reads from *Xenoturbella bocki* branch less then 1 in 1000 vertices due

to sequence variation. The branching of the assembly graph indicates that shorter repeats are more frequent than longer repeats. Frequent repetitive sequences in the genome are difficult to resolve just by using short sequencing reads and often require additional read libraries with a long insert size. Frequent repetitive sequences increase the number of alternative walks in the assembly graph, and make the assembly difficult to resolve. We have sampled the k-mers on the sets of Xenacoelomorpha reads and classified the branches in the de Brujin graph using the preqc src program as repeat branches (Simpson 2014). A high frequency of repeat branches indicates that the Xenacoelomorpha genomes contain multiple repetitive regions (see Figure 2.3).

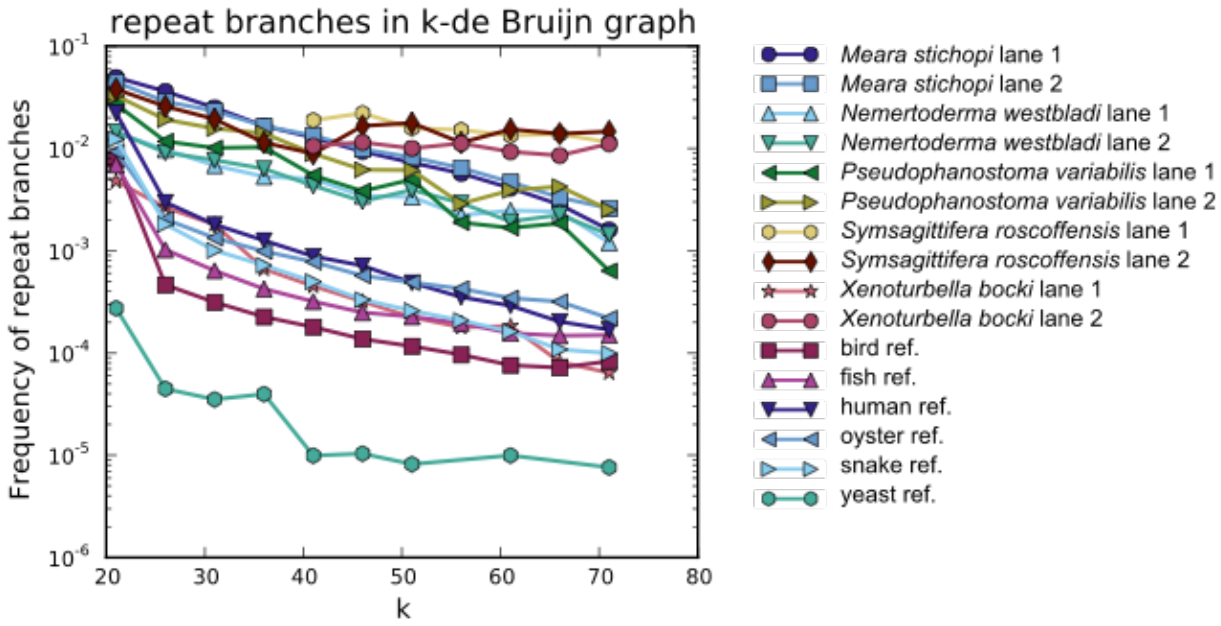


Figure 2.3. A High rate of sequence repeats in the Xenacoelomorpha genomes. Frequency of the de Bruijn graph branching, due to sequence, repeats as a function of k, measured for two sequencing experiments for each Xenacoelomorpha species. The Xenacoelomorpha genomes (top six samples) show a high frequency of the repeat branches compared to the reference sequences from the previously sequenced genomes (bottom five).

2.3.2 Genome size estimation based on read mapping.

We have estimated the genome size using a k-mer distribution (Figure 2.3). However, a k-mer coverage was nearly impossible to estimate, because the k-mer distribution was flat and lacked a characteristic peak, and the estimates provided by that method are therefore practically worthless (The estimations ranged from 2.2 Mbp for *Pseudophanostoma variabilis* up to 449.6 Mbp for *Symsagittifera roscoffensis*). Typically the genome size estimation is based on the position of the peak (mode after the first local minimum in the k-mer distribution, as this peak corresponds to the k-mer coverage, and the genome size can be estimated as: (total *K*-mer number)/(volume peak) (Li et al. 2014)), however in the case of Xenacoelomorpha sequencing data such peak was not present. A flat k-mer distribution means that most k-mers are unique and can be found very few times only (due to high content of heterozygosity), which suggests that the coverage of each haplotype is very low.

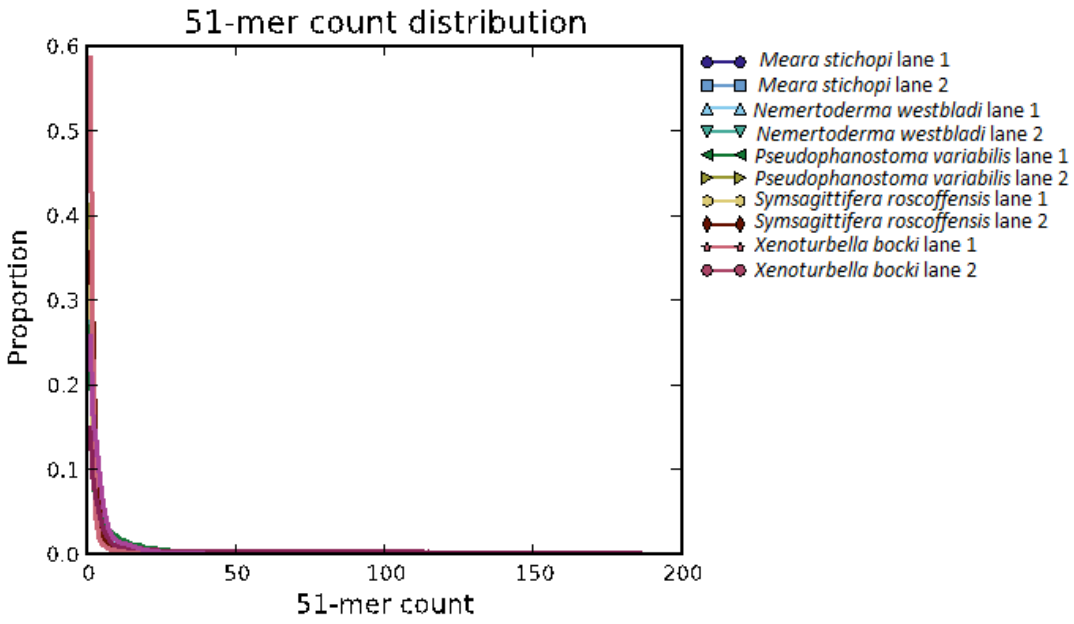


Figure 2.4. A histogram of 51-mer frequencies for each set of Ilumina paired reads from Xenacoelomorpha genome sequencing. The plot lacks second maximum, which indicates genome coverage.

To correct on a previous result, we chose 4 genes, which have a putative single copy in the genome (elongation factor 2, 60S ribosomal protein L18A; 50S ribosomal protein L4) and found them in

the genome assembly of *Symsagittifera roscoffensis, Xenoturbella bocki, Meara stichopi, Nemertoderma westbladi, Pseudophanostoma variabilis, Praesagittifera naikaiensis and Paratomella rubra*. For each genome we chose two 100bp read libraries and mapped the reads to gene sequence. Based on the count of mapped reads we calculated the coverage depth for each pair of the putative single-gene copy ortholog and read library. The Genome size was calculated according to the formula Genome size = (number of reads * read length / 2 * coverage depth).

The estimated genome size of *Symsagittifera roscoffensis, Meara stichopi, Pseudophanostoma variabilis, Paratomella rubra,* and *Xenoturbella bocki* ranges between 0.5 - 1Gbp. This result indicated that appear to be over 3 times smaller than the human genome (3.4Gbp) and similar to the size (to) (of the) previously published *Saccoglossus kowalevskii* 1.1 Gbp genome (Gerhart et al. 2009) and *Strongylocentrotus purpuratus* 0,8 Gbp genome (Sodergren et al. 2006). The genome estimated for *Nemertoderma westbladi* was markedly differed from the estimates of other genome sizes. The estimated genome size of *Nemertoderma westbladi* is 5 to 10 times smaller genome than other closely related Xenacoelomorphs. This may not be that surprising since the genome size varies within Animal Kingdom, with insects ranging between 100Mbp up to 6 Gbp and Amphibians 80Mbp up to 100Gbp (Gregory et al. 2005; Hou et al. 2009). However, strikingly different read converge from the other genome assemblies suggests (49 reads per base for first locus and 63 for the other locus in *Nemertoderma westbladi* for 2 investigated libraries), that the estimation is likely to be inaccurate, and may be a result of a non-uniform coverage distribution. One of the possible explanations for the low coverage could be a high frequency of sequencing errors. However, the error rate per position, measured based on k-mer distribution, for all read sets is lower than 0.05%, and we allowed the read mismatch lower then 0.05%. Another possible explanation of the low coverage could be caused by polymorphism, or the contamination that was not removed by read filtering based on the GC content. In this case, the genome size could be overestimated, because of the islands of heterozygosity.

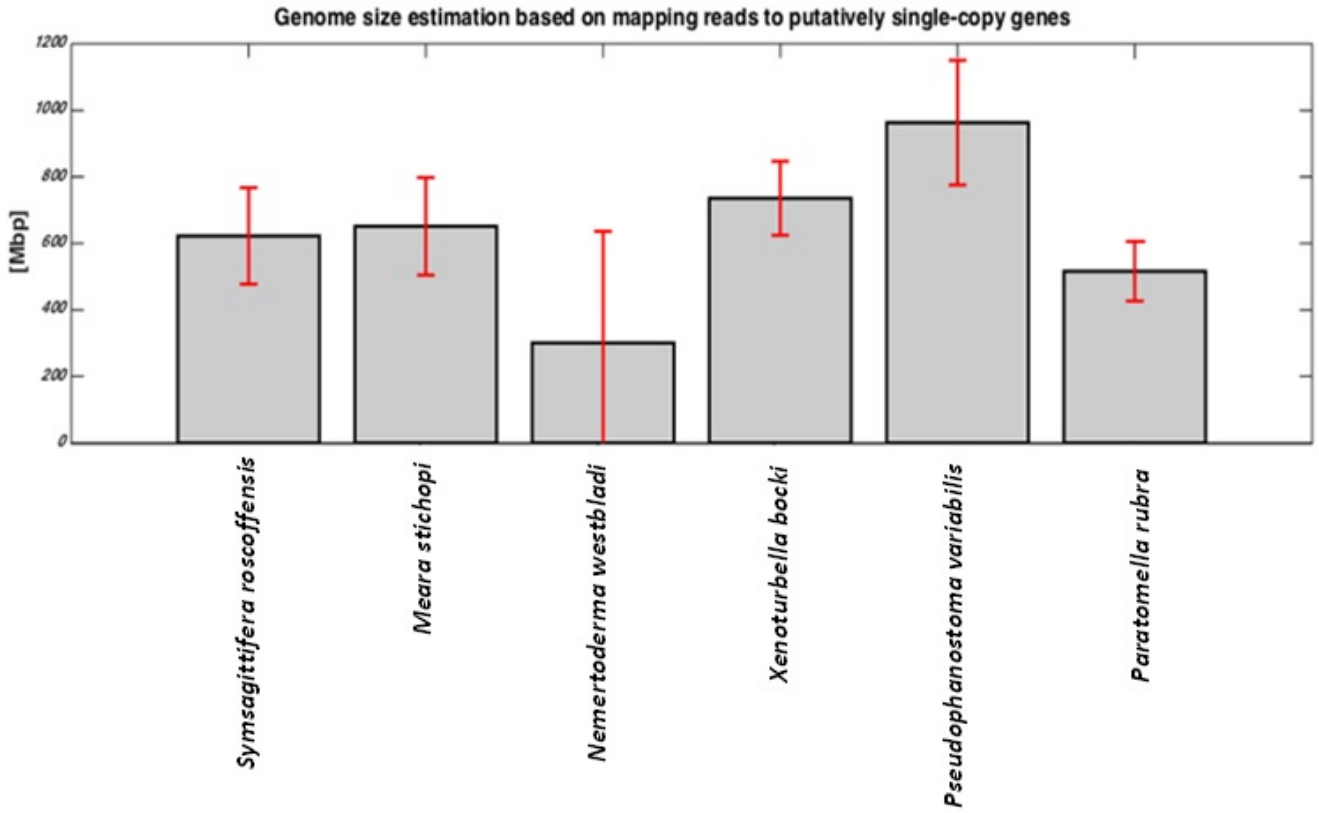


Figure 2.5. Estimated genome size of Xenacoelomorpha species. Genome size for 6 Xenacoelomorpha species was estimated by mapping sequencing reads to single-copy genes elongation factor 2 and RNA polymerase 2.

	genome size estimation	read number	estimated read coverage from the genome size
Symsagittifera roscoffensis	*622.04*	1,317,490,800	106
Pseudophanostoma variabilis	*962.96*	420,359,700	22
Paratomella rubra	*516.06*	583,378,230	57
Meara stichopi	*651.25*	563,402,050	43
Nemertoderma westbladi	*300.37*	724,847,170	121
Xenoturbella bocki	*735.5*	493,800,804	34

Table 2.1 Estimated genome coverage calculated based on genome size estimation and read number.

2.3.3 The assembly quality

2.3.3.1 The N50

6 new genomic assemblies of *Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Xenoturbella bocki, Pseudophanostoma variabilis, Paratomella rubra* genomes were generated from shotgun reads, using the SOAPdenovo2 short-read assembly method (Luo et al. 2012) and the ABySS short-read assembly method (Simpson et al. 2009) (data available thanks to Albert Poustka and Max Telford). We evaluated the assemblies of each genome with the N50 metric (length of the smallest contig out of the minimum set of contigs that cover 50% of the total assembly length) (see Figure 2.5). *Symsagittifera roscoffensis assembly had an N50 of* 2891 bp*, Meara stichopi* 251bp*, Nemertoderma westbladi 361bp, Pseudophanostoma variabilis 886*bp*, Paratomella rubra 722bp Xenoturbella bocki* 10,775bp*.* All the assemblies, apart from *Xenoturbella bocki,* have a very low contiguity, with an N50 metric lower than 3000 base pairs. *Xenoturbella bocki* has an N50 of 10,775 base pairs, which is a similar quality to previously published *Saccoglossus kowalevskii* 1.1 Gb genome with an N50 of 10,074 base pairs (Gerhart et al. 2009) and *Strongylocentrotus purpuratus* 0.8 Gbp genome with an N50 of 13,455 base pairs (Sodergren et al. 2006). However, our values are much lower than other high quality reference genomes of established model organisms (*Drosophila melanogaster* 1.2 Gbp genome N50 of 21,485,538 base pairs; Homo Sapiens 3.4 Gbp genome 56,413,054).

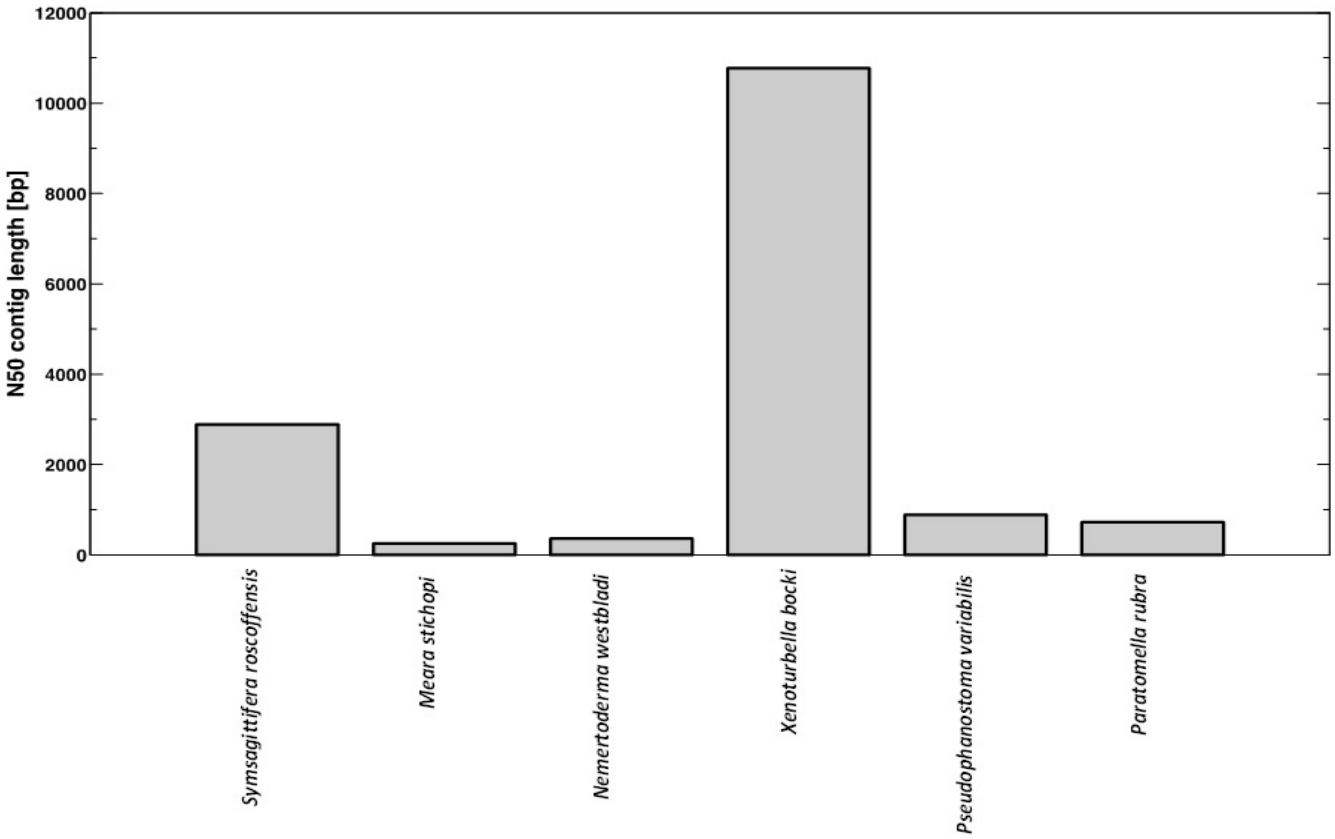


Figure 2.6. N50 contig size of the Xenacoelomorpha assemblies. N50 metric shows low contiguity of *Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Pseudophanostoma variabilis* and *Paratomella rubra*. N50 of *Xenoturbella bocki* is 10,775 base pairs similar value as genomes of Ambulacraria (*Saccoglossus kowalevskii* and *Strongylocentrotus purpuratus*).

2.3.3.2 Number of contigs greater than 10kb

To further evaluate the assembly quality, we have measured a number of contigs in Xenacoelomorpha assemblies that are greater than 10kb. *Xenoturbella bocki* and *Symsagittifera roscoffensis* assemblies contain the most contigs greater than 10kb (long contigs) and are the best in terms of content quality (see Figure 2.6). The few contigs greater then 10kb found in *Meara stichopi, Nemertoderma westbladi, Pseudophanostoma variabilis* and *Paratomella rubra* assemblies highlight potential problems likely to be affecting detection of gene-coding regions. It is worth noticing that we find significantly fewer gene family members in these species, as shown in Chapter 3.

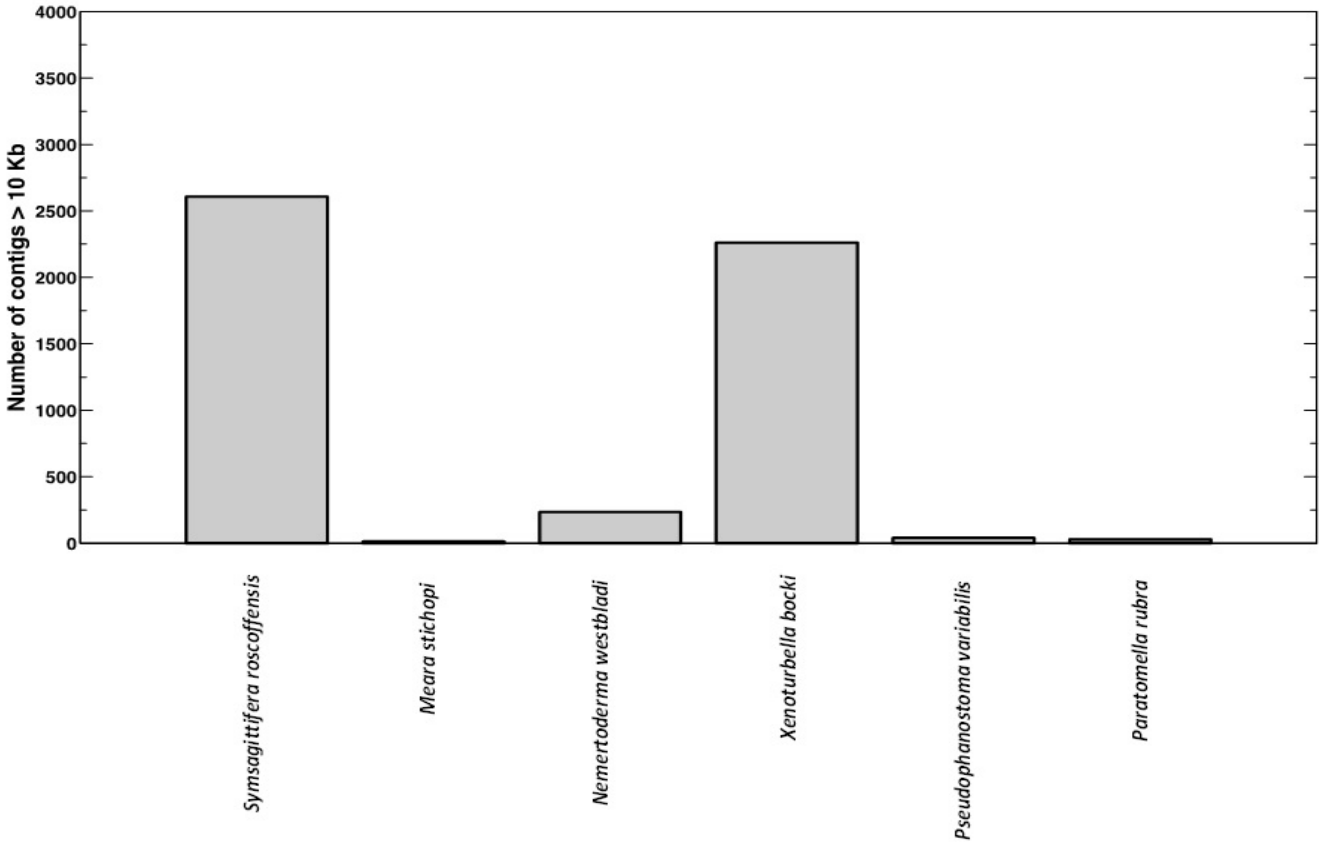


Figure 2.7. *Xenoturbella* and *Symsagetifera* assemblies contain the most contigs greater than 10kb. The number of contigs greater then 10kb in *Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Pseudophanostoma variabilis*, *Paratomella rubra* and *Xenoturbella bocki* assemblies.

2.3.3.3 The N50 scaffold length

Scaffolds are joined contigs, which are based on the information about the distance between paired-end and mate reads. The distance between the pairs of reads is approximate, so the unknown spaces between the mate contigs are filled with an approximate stretch of Ns. To evaluate how well the Xenacoelomorpha assemblies are scaffolded we measured the N50 scaffold statistics (the length of the smallest scaffold out of the minimum set of scaffolds that cover 50% of the total assembly length) (see Figure 2.7). *Symsagittifera roscoffensis assembly had an N50 of* 16,079 bp*, Meara stichopi* 11,411bp*, Nemertoderma westbladi* 9,930*bp, Pseudophanostoma variabilis* 9,492 bp*, Paratomella rubra* 5,619 *bp*

Xenoturbella bocki 110,832 bp. The assemblies of Xenoacoelomorpha can be characterised by low N50 scaffold lengths, below 20Kbp. Other previously published genomes have much longer scaffolds and subsequent N50 scaffold length: *Saccoglossus kowalevskii* 245Kbp (Gerhart et al. 2009) and *Strongylocentrotus purpuratus* 402Kbp (Sodergren et al. 2006), *Ciona intestinalis* 3Mbp (Dehal *et al.* 2002).

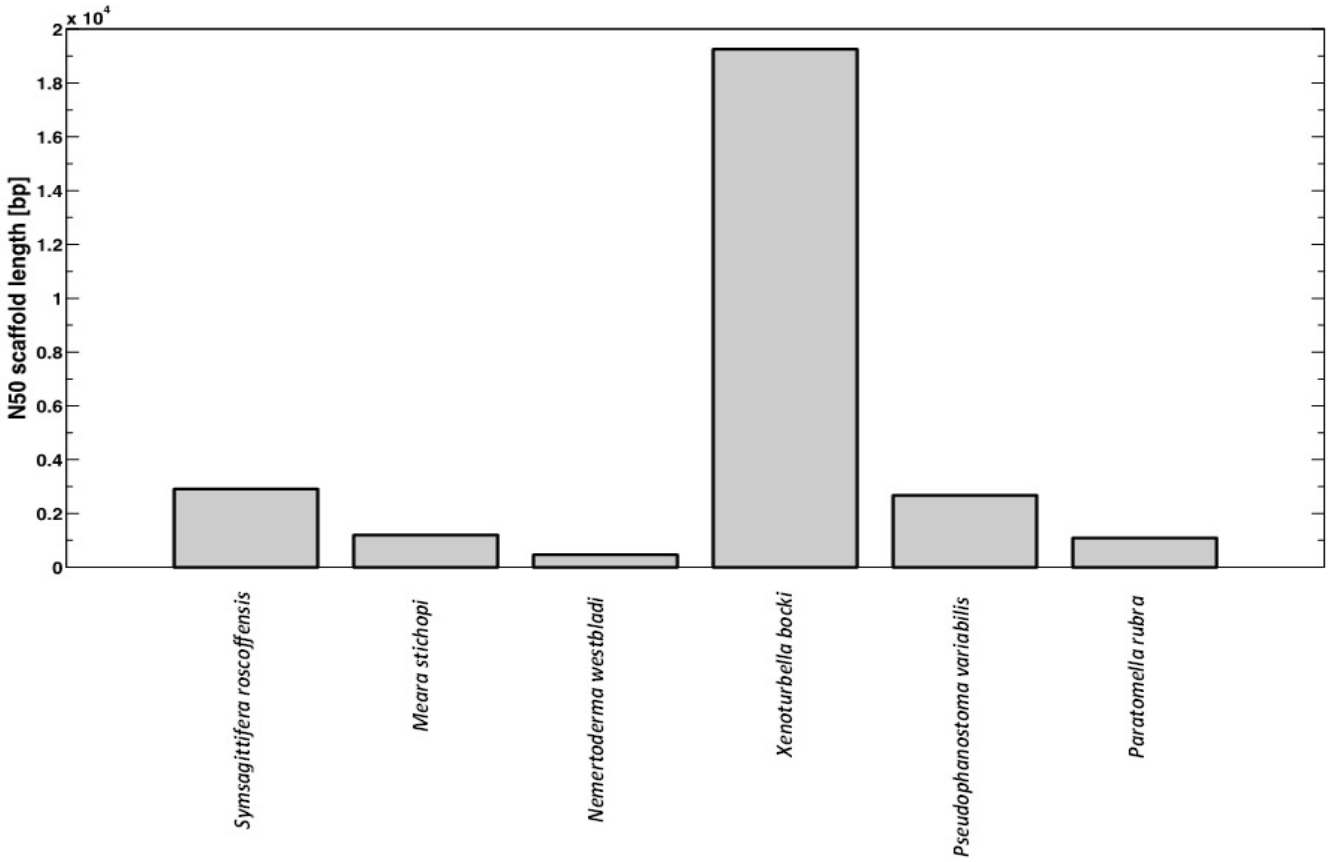


Figure 2.8. N50 scaffold length of the Xenacoelomorpha assemblies. The assembly of Xenoturbella bocki has the mosts long scaffolds over 20000bp and is the best scafolded assembly among all 6 Xenacoelomorpha.

The scaffolds of *Meara stichopi* and *Pseudophanostoma variabilis* contain the highest percentage of gaps in their scaffolds. Nearly 50% of the scaffolds were constructed from unknown sequence, confirming that the assembly failed and suggesting that many parts of the genome are missing from the assembly (see Figure 2.8). This assembly quality statistic directly correlated with the results from Chapter 3, where we investigated the presence of the Deuterostome specific and ancestral metazoan genes in Xenacoelomorpha (see Chapter 3). There, we identified significantly fewer gene family members in *Meara stichopi* and *Pseudophanostoma variabilis* gene predictions from genome assemblies, which contain a large percentage of gaps in the assembly.

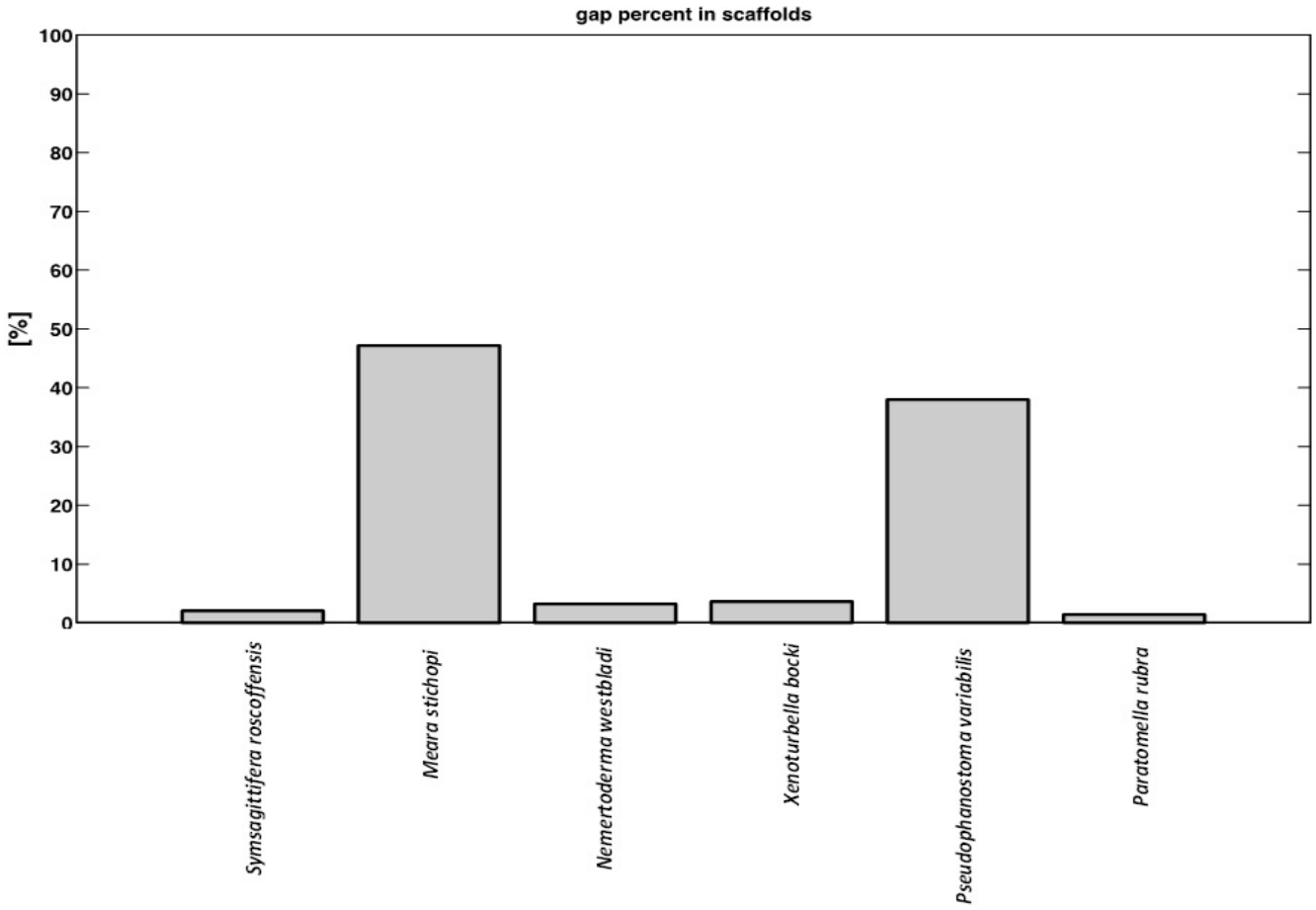


Figure 2.9. *Meara stichopi* and the *Pseudophanostoma variabilis* assemblies contain the highest percentage of gaps in scaffold.

2.3.4 Quality of the Xenacoelomorpha proteomes

We investigated the number of ORFs detected in the transcriptomes of Xenacoelomorpha (see Figure 2.9). In all transcriptome assemblies we detected from 18,000 to 32,000 ORFs, which is around the expected number of genes in most animals. However, it is important to notice that not all of the transcripts are complete (18,495 ORFs total for *Symsagittifera roscoffensis* where 10,988 were complete ORFs, 32,043 ORFs total for *Meara stichopi* where 4,800 were complete ORFs*,* 18,968 ORFs total for *Nemertoderma westbladi where* 2,984 were complete ORFs, 23,209 ORFs total for *Xenoturbella bocki* 9,994 complete ORFs, 22,287 ORFs total for *Pseudophanostoma variabilis* where 4,915 were complete ORFs, 25,703 ORFs total for *Praesagittifera naikaiensis* where 9,226 were complete ORFs*,* and 28,881 ORFs total for *Paratomella rubra* where 13,683 were complete ORFs). Less then one fourth of the transcripts were complete for *Pseudophanostoma variabilis, Meara stichopi* and *Nemertoderma westbladi.* Approximately half of the transcripts of *Paratomella rubra, Praesagittifera naikaiensis, Symsagittifera roscoffensis* and *Xenoturbella* bocki were complete. Furthermore, not all of the genes are expressed in adult animals, so it is very likely that multiple genes are missing from the transcriptome sequences.

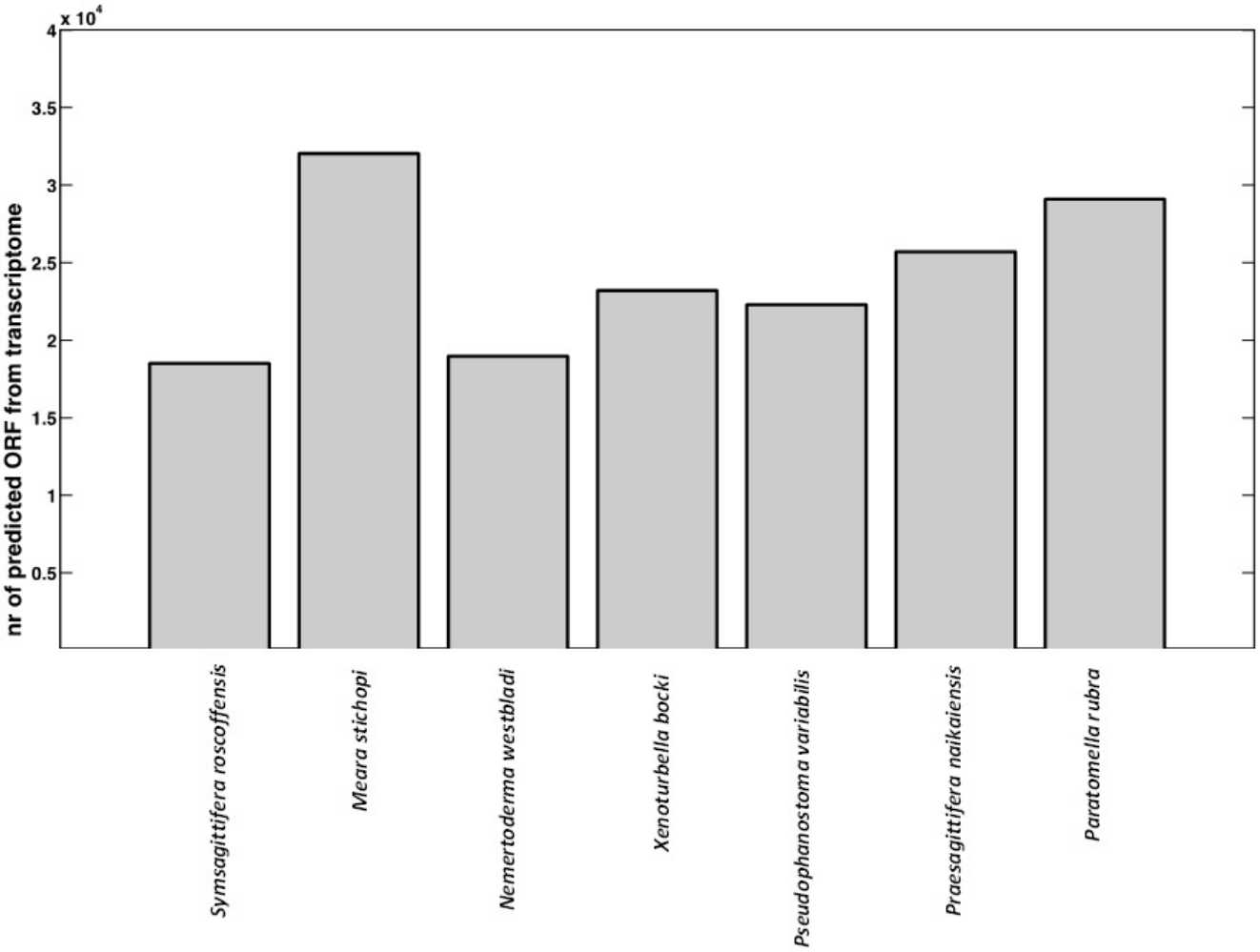


Figure 2.10. Number of predicted ORFs from the Trinity transcriptome assemblies. All 7 Xeneacoelomorpha transcriptomes contain approximetly 20,000 ORFs, which is cloase to an average animal gene number. *Symsagittifera roscoffensis* 10988 complete ORFs*, Meara stichopi* 4,800 complete ORFs*, Nemertoderma westbladi* 2,984 complete ORFs*, Praesagittifera naikaiensis* 9,226 complete ORFs*, Pseudophanostoma variabilis* 4,915 complete ORFs and *Paratomella rubra*. 13,683 complete ORFs, and *Xenoturbella bocki* 9,994 complete ORFs.

To include the genes not present in the transcriptomes of Xenacoelomorpha, we predicted the gene sequences based on a genomic assembly using the GeneScan program (see Methods). For the low quality genome assembly of *Symsagittifera roscoffensis, Meara stichopi, Nemertoderma westbladi, Xenoturbella bocki, Pseudophanostoma variabilis,* and *Paratomella rubra (113,993; 130,115; 80,966; 21,769; 115,245; 52,346* respectively (Table 2.2)), where the N50 contig length was lower than 4000 bp,

over 100,000 genes were predicted for every assembly (see Figure 2.10). However, large number of predicted genes were incomplete (we found 27,388 complete gene predictions in *Symsagittifera roscoffensis, 32,381 in Meara stichopi, 21,156 Nemertoderma westbladi, 26,500 in Pseudophanostoma variabilis,* and *20,053 in Paratomella rubra, 15,126 in Xenoturbella bocki*). It is worth mentioning, that *ab initio* gene predictions are very sensitive (some approaching 100% sensitivity), with a cost of decreased accuracy, as a result of large number of false positives. Even with the fully known sequence of the human genome *ab initio* methods reach 50% accuracy (Guigó et al. 2006). Not all *ab initio* predictions are protein-coding genes, non-protein coding RNA genes, such as miRNAs, snoRNAs and regulatory regions are also recognized by this method.

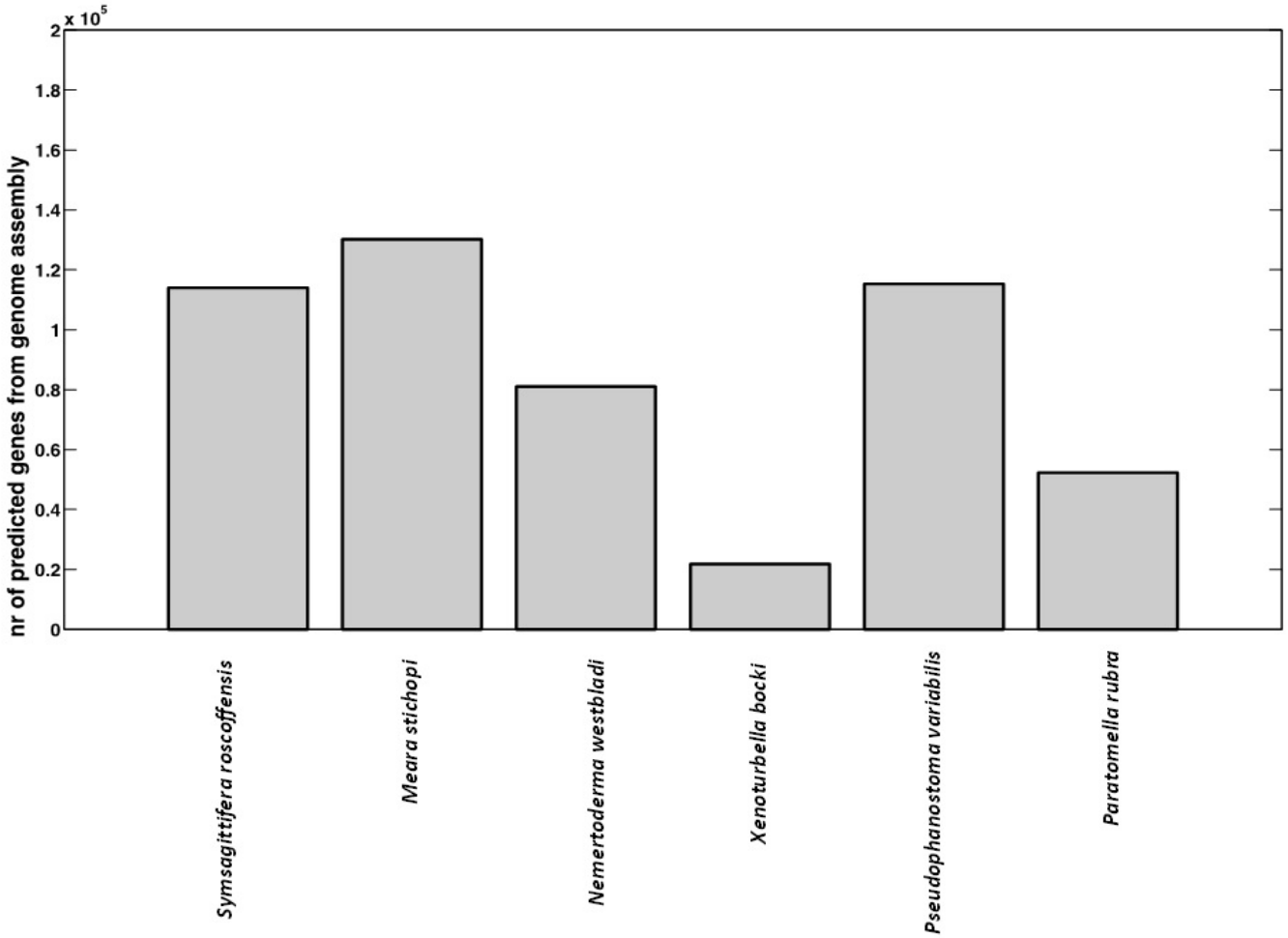


Figure 2.11. Number of predicted genes from the genome assemblies of Xenacoelomorpha. In lower quality assemblies we predict large number of over 100,000, were most of the predictions are incomplete (27,388 complete gene predictions *Symsagittifera roscoffensis, 32,381 in Meara stichopi, 21,156 Pseudophanostoma variabilis (contain start and stop codon at the ends)*).

To obtain most complete protein sets of genes for each species, we joined the predicted genes form both the genome assemblies and the transcriptome assemblies and clustered all the gene predictions using the CD-HIT with a 97% identity threshold (Fu et al. 2012). We removed redundant sequences present in both the transcriptome and the genome predictions, which may differ by several single amino acid changes. We did not join genes that locally align to each other, because it could result

in artificial sequences. At the amino acid level gene sequences can often locally align, because they share a common domain or are closely related paralogs. We predicted 78,346 genes in which 32,456 are complete gene predictions in *Symsagittifera roscoffensis, 125,734* genes in which 35,867 are complete gene predictions in *Meara stichopi, 92,748* genes in which *23,233* are complete gene predictions in *Nemertoderma westbladi, 125,874* genes in which *27,378* are complete gene predictions in *Pseudophanostoma variabilis,* and *44,528* genes in which *24,329* are complete gene predictions in *Paratomella rubra, 31,034 in* genes in which *19,206* are complete gene predictions in *Xenoturbella bocki*. This results shows that the large number of genes predicted from the genome assemblies is not the result of redundant predictions. The number of predicted genes after clustering decreased, but is still much larger than 20,000. This is observed in all the Xenacoelomorpha species apart from *Xenoturbella* (see Figure 2.10). The fact that *Xenoturbella* has also the best quality genome assembly, in terms of contiguity, suggests that protein data set from *Xenoturbella* is the most complete and contain the most full-length gene sequences.

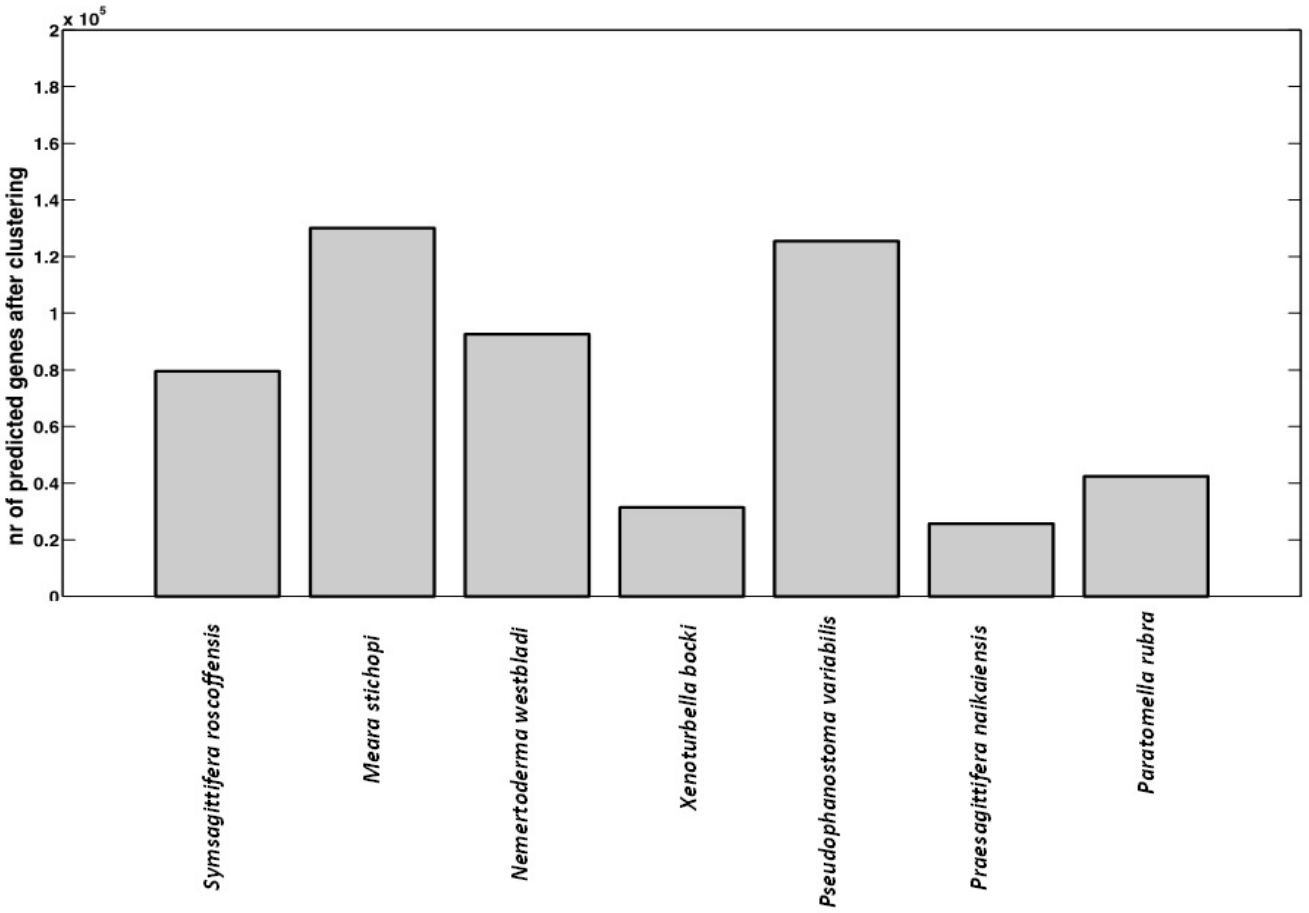


Figure 2.12. Number of predicted genes after clustering with CD-HIT. Redundant sequences were removed and the number of gene predictions from the genome decreased by up to 23%.

	total genomic ORFs	compleat genomic ORFs	total transcriptom ORFs	percentage of unigines with alterative splice variant	average number of isoforms per unigene	compleat transcriptomic ORFs	total number of gene predictions after clustering	compleat ORFs from both transcriptome and genome after clustering
Symsagittifera roscoffensis	113,993	27388	18495	25.27%	1.33	10,988	78,346	32,456
Pseudophanostoma variabilis	115,245	*26500*	22287	28.38%	1.34	4,915	*125,874*	27,378
Paratomella rubra	52,346	*20053*	28881	25.19%	1.28	13,683	*44,528*	*24,329*
Meara stichopi	130,115	*32381*	32043	19.90%	1.24	4,800	*125,734*	35,867
Nemertoderma westbladi	80,966	*21156*	18968	27.04%	1.35	2,984	*92,748*	23,233
Xenoturbella bocki	21,769	*15126*	23209	25.77%	1.36	9,994	*31,034*	*19,206*
Praesagittifera naikaiensis	xxx	xxx	25703	18.20%	1.22	9,226	24,835	8,943

Table 2.1 Number of predicted ORFs for each of the sequenced Xenacoelomorpha species.

2.3.5 The presence of core Eukaryote proteins in animal proteomes

The presence of the core Eukaryote proteins is commonly applied to evaluate the quality and completeness of new animal transcriptomes (Bradnam et al. 2013). Here, we investigated how many of the 100 core Eukaryote genes identified using the OMA standalone 0.99x based on 67 animal proteomes in Chapter 5, are present in each species. We compared the proportion of these core proteins in each of the 67 proteomes, constructed based on transcriptome and reference genome, used in our analysis. We found over 70% of the core Eukaryote proteins in *Symsagittifera roscoffensis, Xenoturbella bocki, Nemertoderma westbladi, Pseudophanostoma variabilis* and *Paratomella rubra* protein sets. A similar level of the core Eukaryote proteins can be found in model chordates such as *Ciona savignyi* or *Branchiostoma floridae.* A low number of 53% of core proteins were present in the *Meara stichopi* proteome, however this value is similar level to proteomes constructed from publicly available trancsriptomes or low quality genomes (Gerhart et al. 2009; Sodergren et al. 2006; Dehal *et al.* 2002).

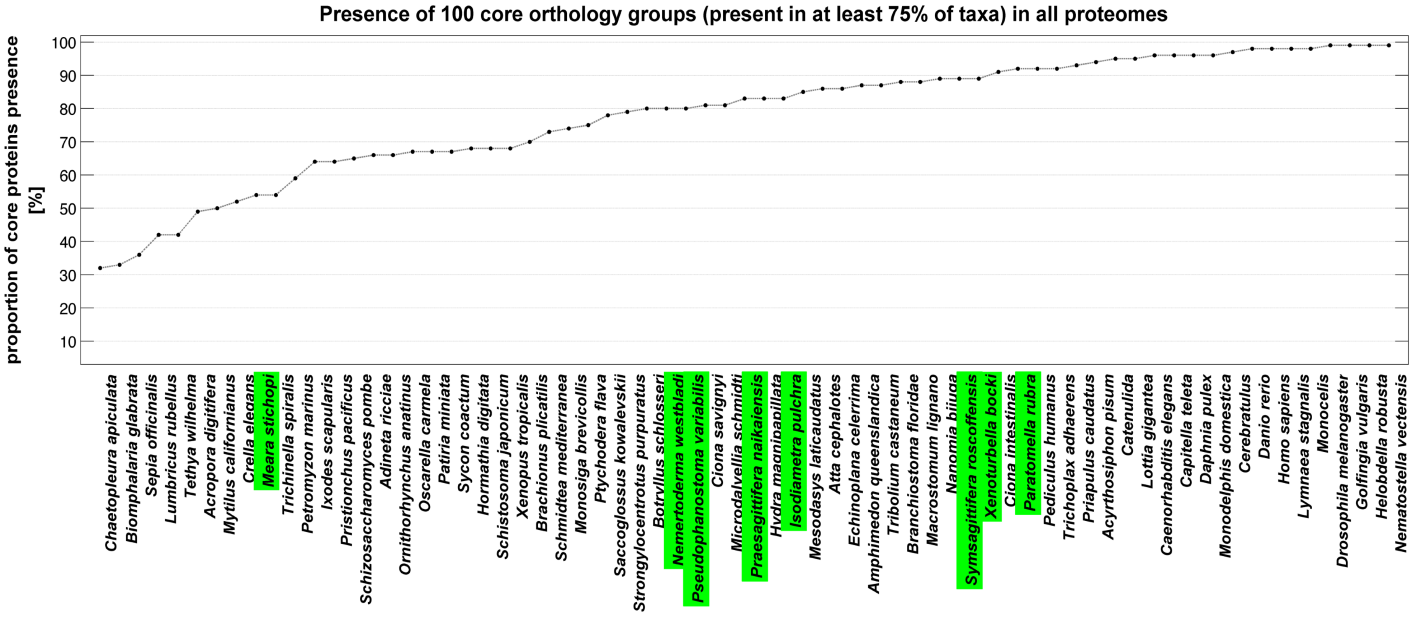


Figure 2.13. The presence of the core Eukaryotic proteins in animal proteomes. Acoelomorphs indicated in green.

2.3.5.1 The use of multiple proteomes improves the ability to find core Eukaryote genes within the Xenacoelomorpha clade

Finding members of an orthology group within a clade is essential for the analysis of the inference of the ancestral animal genomes content and the reconstruction of the evolutionary events within the animal kingdom, as presented in Chapter 4. We examined whether the use of multiple genomes improves our ability to find core proteins within a clade. We randomly generated a subset of the Xenacoelomorpha proteomes and investigated the presence of the core Eukaryote genes in at least one of the proteomes (see Figure 2.12). We found, not surprisingly, that the ability to find core Eukaryote genes in the subset of the Xenacoelomorpha proteomes improves with the number of proteomes used in the analysis. Therefore, we improved the analysis of the gene content in the Last Common Ancestor of Xenacoelomorpha by using multiple genomes in our analysis. Based on this result we expected to be able to infer the content of the genome of Last Common Ancestor of Xenacoelomorpha, even though the inference of the gene gain and losses on the terminal branches of animal tree is probably heavily reliable on the completeness of the protein sets. This also indicates that we minimized the impact of missing sequencing data on gene content analysis of the Xenacoelomorpha clade, by the use of multiple proteomes.

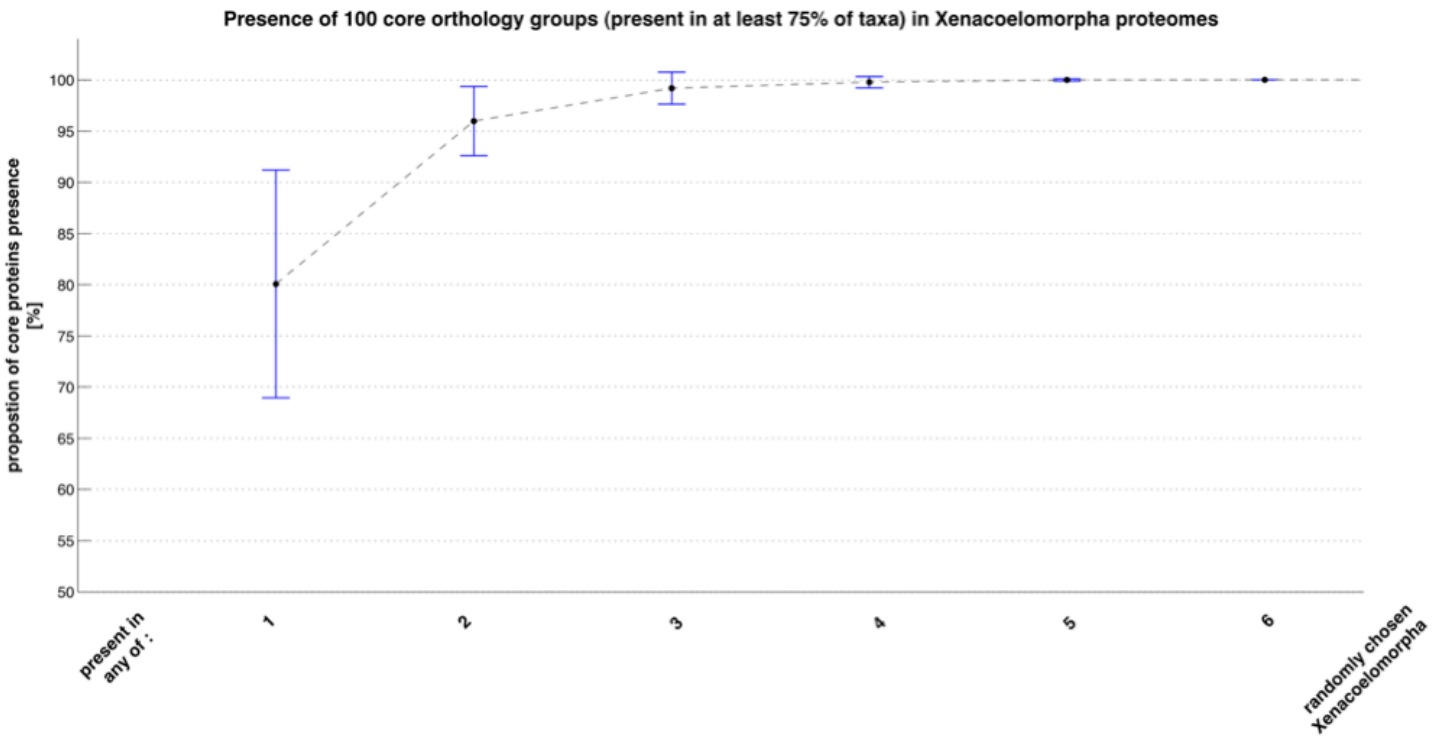


Figure 2.14. Proportion of the present core Eukaryote gene in the subset of the Xenacoelomorpha proteomes improves with the number of proteomes used in the analysis. We randomly chose the subset of 1,2,3,4,5 and 6 genomes and measured the proportion of core gens found in the subset. Error bars represent the standard deviation from the average value.

2.3.5.2 The use of multiple proteomes also improves the ability recognize frequently present genes within the Xenacoelomorpha clade

Similar to the test presented above, we examined if the use of multiple genomes improves the ability to find frequently present, present in at least 50% of species, genes within the clade (collected based on OMA orthology groups from 57 Metazoa species (see chapter 5)). We randomly generated a subset of the Xenacoelomorpha proteomes and investigated the presence of frequently present Eukaryote genes in at least one of the Xenacoelomorpha proteomes (see Figure 2.14). We found that it is easier to find a gene in the subset of the Xenacoelomorpha proteomes, the bigger the subset is. Surprisingly, more than 99% of the frequently present proteins (present in 50% of the taxa) can be found in at least one of 6 randomly chosen Xenacoelomorpha proteomes.

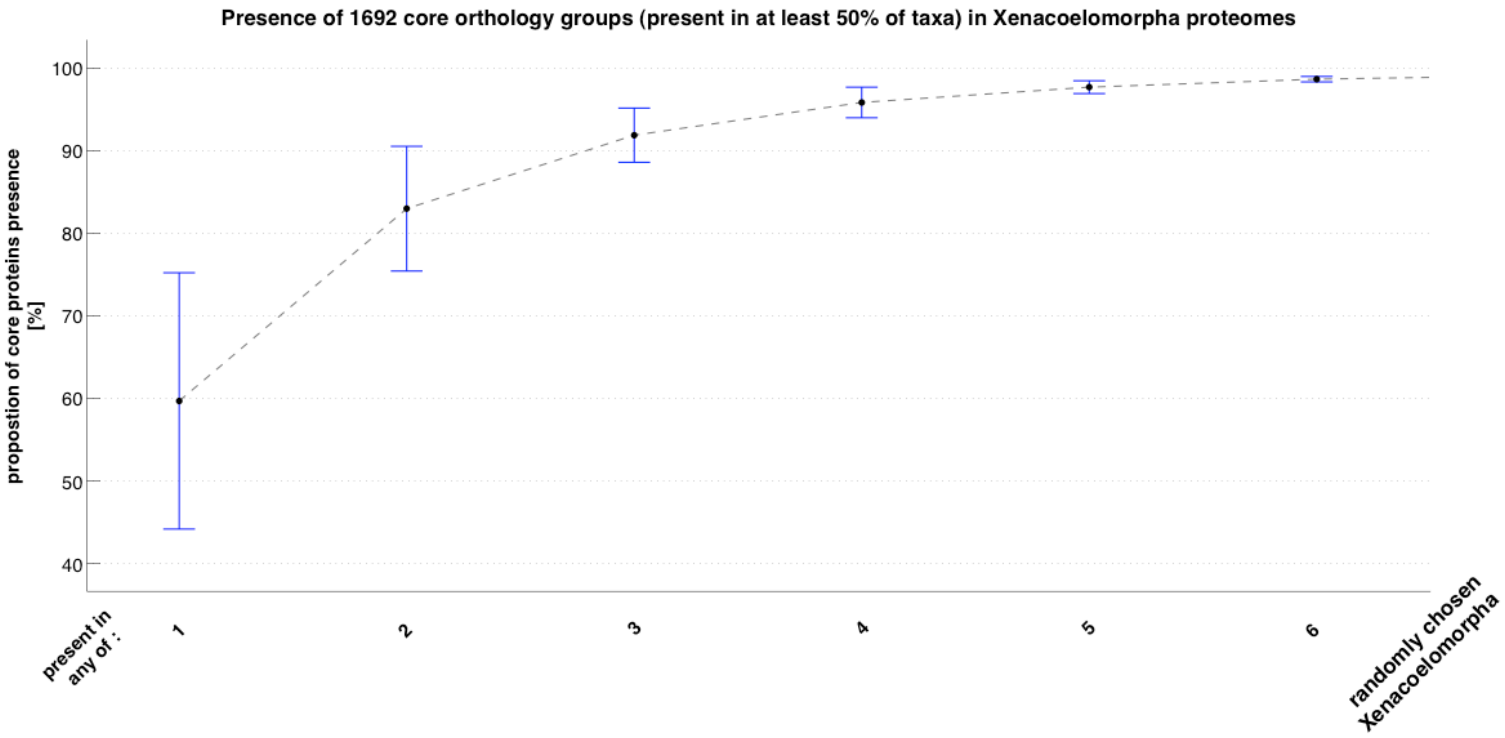


Figure 2.15. Proportion of present frequently present Eukaryote gene (present in at least 50% of taxa) in the subset of the Xenacoelomorpha proteomes improves with the number of proteomes used in the analysis.

2.3.6 Detecting contamination in the predicted protein datasets

After the decontamination with PhymmBL, we assessed that the assembly contained high level of the heterozygosity using the sga preqc program. Based on this result and the shape of k-mer distribution, which didn't show a characteristic peak, we suspected that the contamination might be still present in the datasets. Lately, we have identified contamination in protein predictions of Xenacelomorpha using alternative Kraken method (Wood et al. 2014; as described in section 2.2.6)(cooperation with Jean-François Flot). A low proportion of genomic protein predictions contained contamination from other species. However, a large number of *Meara stichopi* (41.3%) and *Xenoturbella bocki* (42.5%) sequences in the transcriptomic protein predictions is of a contaminant (Figue 2.16). Most of the contamination in *Xenoturbella bocki* came from human (21.6% of all sequences), while most of the contamination in *Meara*

stichopi came from pathogen aerobic Protobacteria *Burkholderia gladioli (16.7%* of all sequences). In the other transcriptomes, the inferred contamination was considerably lower (see Fig. 2.16). We further addressed the possibility of contamination in further analysis showed in this thesis. In the analysis of clade specific gene families presented in Chapter 3, I checked for identical matches in other species in NCBI using BLAST and made a phylogenetic tree for each family (the human contamination would be in that case noticeable on a tree, as the branch connecting human and Xenacoelomorpha protein would be extremely short or equal to 0). We have removed the contamination from the OMA orthology groups (in a cooperation with Herve Philippe) by removing identical matches from other Metazoa species using BLAST before using it in the further phylogenetic analysis in Chapter 6.

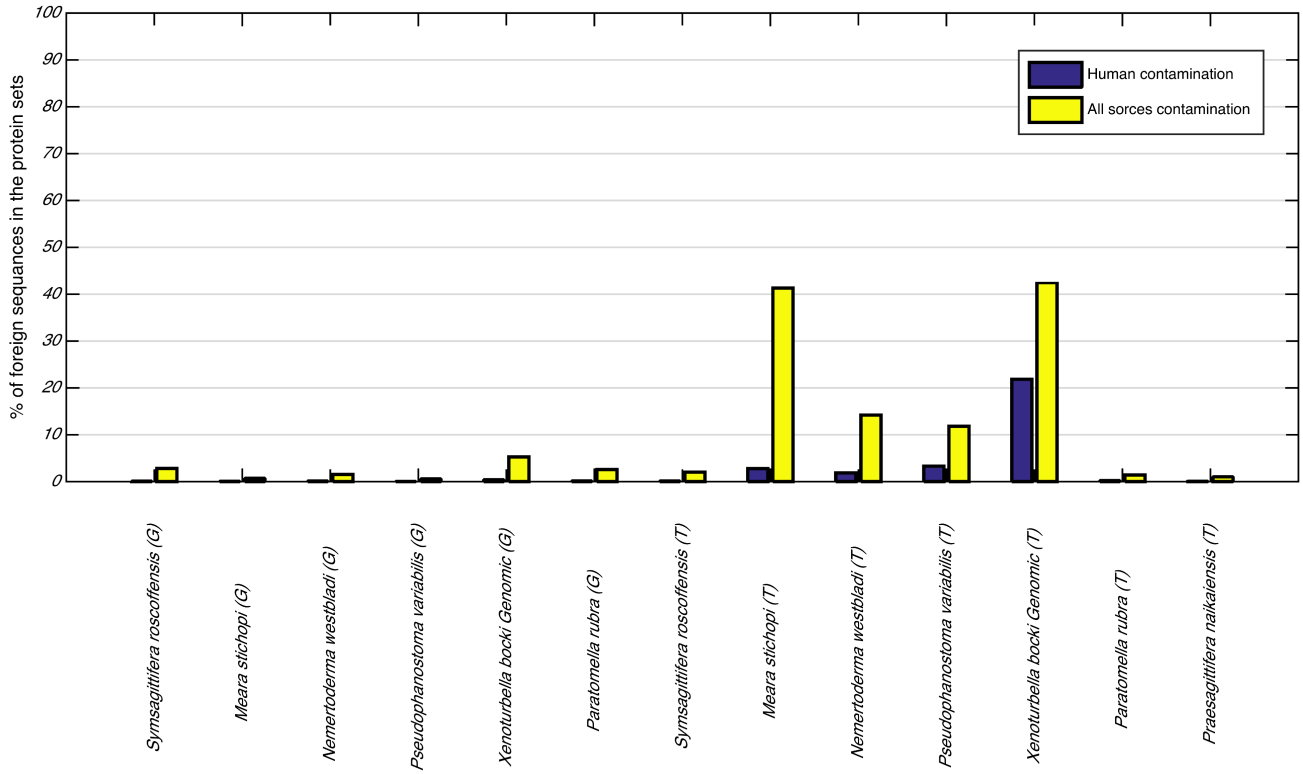


Figure 2.16. Proportion of contaminated sequences coming from human (blue), and all sources identified by Kraken software in Xenacoelomorpha protein predictions.

2.3.7 Discussion

Sequencing reads of Xenacoelomorpha genomes are rich in sequence variation, repeats and have a low genome coverage, which resulted in poor quality genome *de novo* assemblies. *Pseudophanostoma variabilis, Meara stichopi Nemertoderma westbladi* genome assemblies can be characterize by poor contiguity, high percentage of unknown genome and high number of incomplete gene predictions. Poor data quality will result in lower recall of the gene family members in these species using BLAST methods where incomplete genes will lower the local alignment score (see Chapter 3). However, we have complemented the data with the transcriptome assemblies in order to obtain most complete protein sets for 7 Xenacoelomorpha species, which should result in the improvement of the

completeness of the protein sets (as false positive gene predictions were removed from the analysis in later stages in chapter 5). The content of core proteins in Xenacoelomorpha, inferred by OMA standalone algorithm using multidirectional similarity search, is high and does not differ much from other reference organisms, indicating good completeness of Xenacoelomorpha protein sets. Additionally, the use of multiple protein sets of Xenacoelomorpha at once improves the recognition of the core proteins in the clade even further, which is encouraging and minimizes the effect of missing data in future analysis of ancestral genome content inference and animal phylogeny reconstruction (see Chapter 5,6).

Chapter 3

Analysis of gene family content in Xenacoelomorpha genomes using PhylomeDB database

3.1 Introduction

The two possible locations for the Xenacoelomorpha within the animal tree are i) as the sister clade to all other bilaterians (Wallberg et al. 2007; Jondelius et al. 2002; Littlewood et al. 2001; Ruiz-Trillo et al. 2002; Telford et al. 2000; Telford et al. 2003; Hejnol et al. 2009; Srivastava et al. 2014) and ii) as deuterostomes, most closely related to the Ambulacraria (echinoderms and hemichordates) (Nakano et al. 2013; Bourlat et al. 2006; Telford et al. 2008; Philippe et al. 2009, 2011).

A position outside of the Bilateria fits well with the relative simplicity of these worms, both morphological (for example they lack a through gut and most organs) and genetic (they have a maximum of 4 of the 8 Hox genes typical of Bilateria and lack a number of bilaterian microRNAs) (Fritzsch et al. 2007). Genetic simplification of gene content is likely to be associated with the apparent morphological simplicity of the worms, and can explain the lack of characteristics typical for other bilaterians. However, if we consider phylogenetic position of xenacoelomorphs within deuterostomes (Philippe et al. 2011) they should possess (or have modified or lost) characteristics of this specific to deuterostomes, which include features like radial cleav-age, deuterostomy, enterocoely, gill slits, endostyle, and postanal tail (Gerhart et al. 2005). There are no coelomic cavities, no anusor signs of gill slits in simple worms. On the other hand, it has been shown that a number of the miRNAs missing from acoelomorphs have been lost rather than being primitively absent (they are retained by *Xenoturbella*) and several deuterostome specific genes, including the sole known deuterostome specific miRNA (miR-103), are also found in xenacoelomorphs (Philippe et al. 2011). There is in addition a single miR-2012 that is specific for Ambulacraria and Xenacoelomorpha. Some genetic characteristics specific to deuterostomes

may be still present in Xenacoelomorpha genome. Genetic correlates that are only present in deuterostomes and no other taxa so far sequanced, will be informative for the phylogenetic position of the worms. One such gene (gene Rsb66) is only found in deuterostomes and Xenacoelomorpha. We aim to extend previous reports about the gene loss in Xenacoelomorpha by investigating their gene content to try to find the evidence for degenerative evolution.

3.1.1 Gene families as clade specific characteristics

We were interested in extending previous limited *ad hoc* analysis and try to find deuterostome specific genetic characteristics in Xenacoelomorpha. To analysed gene content of Xenacoelomorpha we have used a number of genomic and transcriptomic resources from *Xenoturbella bockii*, 5 species of acoel and 1 species of nemertodermatid to look for the presence or absence of genes and gene families present in the deuterostomes and more generally within the Bilateria. We were interested in whether we could detect, on the one hand, a high level of absence of the bilaterian genes/gene families within the xenacoelomorphs as might be expected if they branched before the protostome/deuterostome divergence and, on the other hand, we have looked for the presence of the gene families linking xenacoelomorphs specifically to the deuterostomes as predicted from a deuterostome affinity.

To achieve this, we have taken as a starting point the gene families curated within the phylomeDB database. “Gene family” is a term generically used to describe a collection of genes or proteins that are presumed to share common ancestry (Henikoff et al. 1997). A gene family is a set of several homologous genes, formed by duplication of a single original gene. Genes within gene families evolve through duplication, conversion and speciation events (Ohta et al. 1991).

3.1.2 PhylomeDB database as a resource in gene family search

PhylomeDB is a publicly available repository of complete phylomes that allows researchers to access and store large-scale phylogenomic analyses. PhylomeDB is a database of phylomes built on a complete set of proteins from model organisms like human, the yeast *Saccharomyces cerevisiae* and

the bacterium *Escherichia coli* and several others. All phylomes in the database are built using a high-quality phylogenetic pipeline that includes evolutionary model testing and alignment trimming phases. For each genome, PhylomeDB provides the alignments, phylogenetic trees and tree-based orthology predictions.

For each of these proteins, the sets of homologs proteins from a defined set of related organisms are found using Smith–Waterman algorithm (Smith et al. 1981). The search is performed against the corresponding proteome dataset to retrieve a set of proteins with a significant similarity (*e*-value $<10^{-3}$). For each protein family ML trees are reconstructed with four different evolutionary models (JTT, WAG, BLOSUM62 and VT). Next, the best ML tree is determined by finding best fitted model with the Akaike Information Criterion (AIC). Bayesian tree for each gene family is calculated with MrBayes using best fitting ML tree as a starting topology. A reference tree produced by this Bayesian reconstruction is a consensus phylogeny however, partitions with a posterior probability lower than 0.5 are collapsed.

The orthology predictions of gene families are generated for each seed sequence by mapping duplication and speciation events on the reference tree. A species-overlap algorithm annotates nodes on the gene tree (Huerta-Cepas et al. 2007). The algorithm checks all nodes that connect the seed protein to the root of the tree and marks it as a duplication event, if its two children nodes share one or more species. Using the phylomeDB resource, we were able to consider all gene families present in the human genome and, through interrogation of the database, identify the patterns of presence and absence of each of these families in other metazoan clades and in several out group taxa. In this way we were able to identify a set of genes (gene families) that were restricted either to the chordates, to the deuterostomes or to the Bilateria.

We next tested, using BLAST searches of much more complete online nucleic acid and protein databases, that the genes identified in this way really were restricted to the clades mentioned rather than simply being absent from the taxa represented in phylomeDB. Finally, we asked how many of the bilaterian genes were absent from the Xenacoelomorphs and how many of the deuterostome specific

genes were present in the Xenacoelomorphs and whether, compared to other taxa known to be outside the Bilateria or within the deuterostomes, these numbers were suggestive of support for either of the contested phylogenetic positions for the Xenacoelomorphs.

3.2 Materials and methods

We identified the presence of the clade specific gene family members in Xenacoelomorpha protein sets using computational pipeline, which consisted of following steps:

1. Identifying the candidates for clade specific gene families in the PhylomeDB database using some parses_PhyDB.pl software we developed
2. Verifying the absence of similar homologs in outgroup species using in-depth BLAST similarity search
3. Preparing non-redundant protein sets from Xenacoelomorpha genome and transcriptome assemblies
4. Testing for presence of the verified clade specific family members using family-RBH (Reciprocal Best Hit) algorithm

3.2.1 Identifying clade specific genes/gene families' candidates using PhylomeDB databse

Our starting point was the phylomes available in PhylomeDB (Huerta-Cepas et al. 2014). Phylomes are based around the complete protein set from a given seed organism (we used Human phylome as the taxonomic range of orthology predictions was best suited to this analysis). The full set of predicted orthology and paralogy relationships for each phylome, which was calculated based on the relations within human phylome based on reference Bayesian trees using species-overlap algorithm with MetaPhOrs software (Pryszcz et al. 2011), was downloaded from the download section of the database.

We grouped the genes from all Metazoa species that have one-to-one, one-to-many and many-to-many predicted relation to human seed protein into gene families. Next, we identified human gene/gene families whose orthologs/co-orthologs in other taxa had specific restricted taxonomic distribution patterns: present only in chordates; present only in the deuterostomes or present only in the bilaterians using parses_PhyDB.pl software we developed. To be considered a clade specific gene family, we required that they are present in members of both child branches of the clade's last common ancestor, which implies that the gene family was present in the last common ancestor of the clade, and that there are no orthologs present in out-group species (this is further verified in verification step (see Section 2.2)).

To achieve this, a perl script that parses the gene content in each of the reference species for each gene family, was written. The program analyses the pattern of the presence and absence of genes in taxonomic clades (Metazoa, Bilateria, Deuterostomia, Ambulacraria, Chordata, Vertebrata, non-vertebrate, Chordata).

As alluded to above, we defined three groups of clade specific gene families for the purposes of this analysis:

- A gene family was classified as specific to Bilateria if at least two members of the predicted PhylomeDB family can be found in Deuterostomia, at least two members of the family can be found in predicted PhylomeDB Protostomia and no member could be found in the outgroup (all non bilaterian species (Arabidopsis thaliana, Chlamydomonas reinhardtii, Saccharomyces cerevisiae, Yarrowia lipolytica, Debaryomyces hansenii, Neurospora crassa, Cryptococcus neoformans, Candida albicans, Candida glabrata, Gibberella zeae, Leishmania major, Plasmodium falciparum, Paramecium tetraurelia, Encephalitozoon cuniculi, Kluyveromyces lactis, Dictyostelium discoideum, Dictyostelium discoideum, Guillardia theta, Plasmodium yoelii, Ashbya gossypii))
- A gene family was classified as specific to Deuterostomia, if at least two members of the predicted PhylomeDB family could be found in Chordata, at least one members of the predicted PhylomeDB

family could be found in Ambulacraria (due to the low number of sequenced genomes in Ambulacraria) and no member could be found in the outgroup (all non deuterostomes species)

- A gene family was classified as specific to Chordata, if at least two members of the predicted PhylomeDB family could be found in a group of Vertebrata, at least two members of the predicted PhylomeDB family can be found in a group of non-vertebrate chordates (Cephalochordata or Urochordata) and no member could be found in the outgroup (non chordate deuterostomes and non deuterostomes)

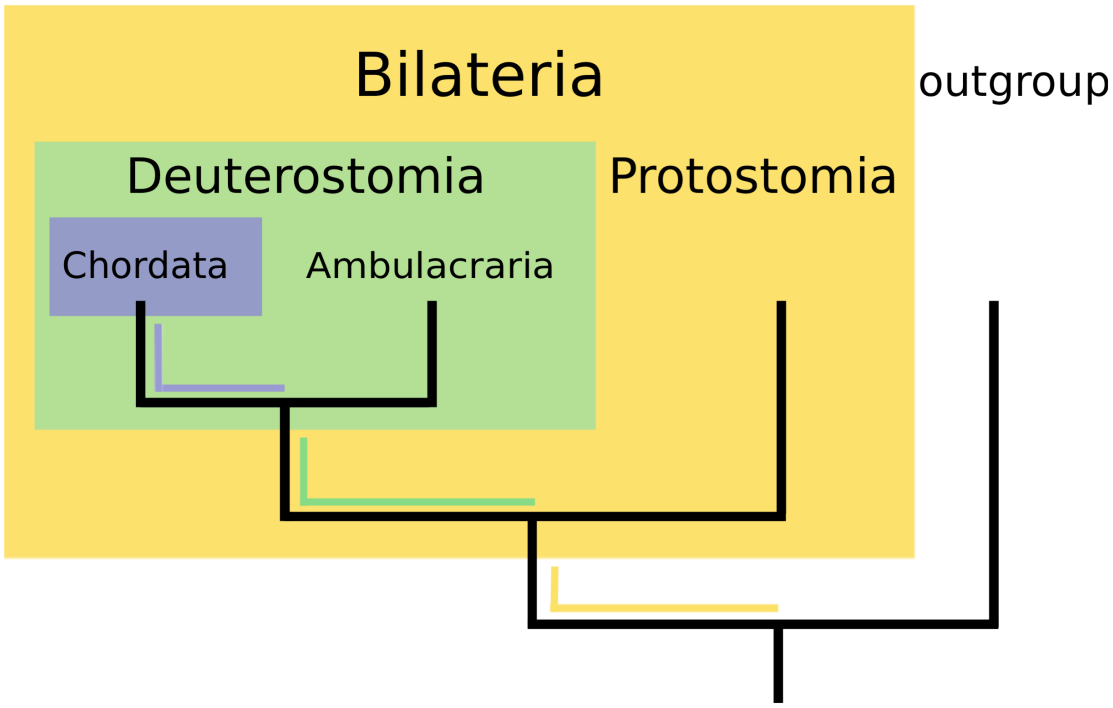


Figure 3.1. A schematic representation of Chordata, Deuterostomia, Bilateria clade specific gene families. A blue box represents Chordata specific gene families, obtained after the speciation of Ambulacraria and Chordata on a branch leading to Chordata (indicated with blue line, these genes are present in both Vertebrata and non-vertebrate chordates but absent in Ambulacraria, Protostomia and the outgroup species). Green box represents Deuterostomia specific gene families, obtained after the speciation of Deuterostomia and Protostomia on a branch leading to Deuterostomia (indicated with green line, these genes are present in both Chordata and Ambulacraria but absent in Protostomia and the outgroup species). Yellow box represents Bilateria specific gene families, obtained after the speciation of Bilateria from non-Bilateria metazoans on a branch leading to Bilateria (indicated with yellow line, these genes are present in both Deuterostomia and Protostomia but absent in the outgroup species).

3.2.2 Verifying absence of putative clade specific genes/gene families using BLAST searching of online databases

Once we had identified a candidate for gene family in PhylomeDB that appeared to be restricted to a clade of interest, we wished to test for the presence of the character in additional species that are not included in PhylomeDB database. PhylomeDB is limited to genomes from 715 species of which just 31 are metazoans. To confirm the absence of similar homologs in the out-group species we performed a multiple BLAST searches using the following NCBI resources: nr, est, RefSeq, wgs and htgs databases, expanding the analysis to as many as 221,263 Eukaryotic species.

3.2.2.1 BLAST search settings

We performed BLAST similarity search using human protein sequences as a query with the blastp program against the protein database and the tblastn program against the nucleotide database (the type of the algorithm depends on the format of target databse (nr, est, RefSeq, wgs and htgs databases were used)). The search was restricted to a specific taxonomic range using appropriate condition (E.g. predicted Chordata specific gene families were searched for similar sequences outside Chordata using condition “Eucariota NOT Chordata”). Candidate clade specific gene families were discarded if we found a protein with a significant BLAST hits (E-value < 10^{-3}) in any outgroup to the clade in question.

3.2.2.2 NCBI resources used for in depth BLAST similarity searches

- nr (non-redundant sequence database) – contains non-redundant sequences translations from GenBank, PDB, SwissProt, PIR and PRF
- refseq (The Reference Sequence) – curated non-redundant sequence database of genomes
- est - contains short single-pass reads of cDNA (transcript) sequences

- wgs (Whole Genome Shotgun) – contains genome assemblies of incomplete genomes or incomplete chromosomes of prokaryotes or eukaryotes that are generally being sequenced by a whole genome shotgun strategy
- htgs (The High Throughput Genomic) – contains contigs greater than 2 kb from genomic sequence projects which are made available to the scientific community before their publication

3.2.3 Xenacoelomorph sequence data

3.2.3.1 Genomes and transcriptome assembly

Genomic DNA from *Xenoturbella bocki*, from the nemertodermatids *Meara stichopi* and *Nemertoderma westbladi* and from the acoels *Symsagittifera roscoffensis, Pseudaphanostoma variabilis* and *Paratomella rubra* were sequenced using Ilumina technology, producing 100nt paired end reads. Reads were assembled using SOAPdenovo2 (Luo et al. 2012) and by Albert Poustka (see Chapter 1). The transcriptomes of all six xenacoelomorph species were assembles from Illumina RNA-seq reads with SOAPdenovo-Trans by Albert Poustka (see Chapter 1).

3.2.3.2 Xenacoelomorph data: protein prediction

Core protein genes were identified in genome assemblies using CEGMA (v2.4) (Parra et al. 2007). Second set of genes was indentified in the genome assembly using PASA (Haas et al. 2003) by mapping assembled RNA-Seq data. Both gene sets were merged (non-redundantly) and used as the training set to optimize AUGUSTUS parameters (Stanke et al. 2008). AUGUSTUS was run on the genome assemblies using the optimized parameter set. Genomic were translated into all possible open reading frames (ORFs) larger than 20 amino acids using custom Perl scripts (by Albert Poustka).

Coding regions in the transcriptome assemblies were detected using TransDecoder software (Haas 2013). Both genomic and transcriptomic gene predictions were incorporated into a single dataset

for each species. Redundant gene predictions were removed from the dataset using USEARCH global alignment clustering with 97% identity (Edgar 2010).

3.2.4 Testing for presence of a family members using family-RBH (Reciprocal Best Hit) optimising an appropriate e-value cutoff and p-value tolerance

3.2.4.1 The family-RBH algorithm for identifying members of the same gene family

In this study, we identified members of known gene families in the Xenacoelomorpha protein sets created from unannotated transcriptome and genome assemblies using an implementation of Reciprocal Best Hit algorithm (family-RBH we developed), similar to the previously described approaches (Fulton DL et al. 2006). This approach, different from classical Reciprocal Best Hit algorithm (RBH) allowed us to identify members of the same family that are not necessarily orthologs. Similar to RBH fist performs BLAST search in the target proteome using one of the family members as a query (human gene). Next, the algorithm performed the revers BLAST searches in the human genome, using top hit from the first search and all the other top hits that are within p-value tolerance coefficient away from the top hit, as a query. Our algorithm classified a gene as member of a family, if any of the top hits to the query within p-value tolerance (the allowed similarity score as a proportion of the best hit) and BLAST similarity scores threshold (e-value cutoff) recognised the same family as the top BLAST hit in the reverse BLAST search.

3.2.4.2 Family-RBH testing, Parameter fitting, optimizing false positive and false negative gene classification

We fitted the parameters of the algorithm to achieve the best positive hits ratio, while keeping minimum number of false positives. We created a training dataset for which we have the information about family membership. We randomly chose subsets of gene families present in *S.cerevisiae, N.vectensis* and *C.elegans* and the subset of gene families (equal size) that do not have family members in *S.cerevisiae, N.vectensis* and *C.elegans*. We ran the family-RBH algorithm using human seed proteins as the starting query and *S.cerevisiae, N.vecte*nsis and *C.elegans* as a primary BLAST database and human genome as a secondary database, with different E-value threshold and p-value tolerance coefficient. Next, we calculated positive to negative results ratio for each pair of the E-value threshold

(in the range of 10^{-1} to 10^{-13}) and the p-value tolerance coefficient (in the range of 0.75 to 1). For the further use of family-RBH algorithm in this chapter, we chose E-value threshold 10^{-3} and p-value tolerance coefficient 0.85, which give 100% of positive family member recognition and 10% of negative family member recognition.

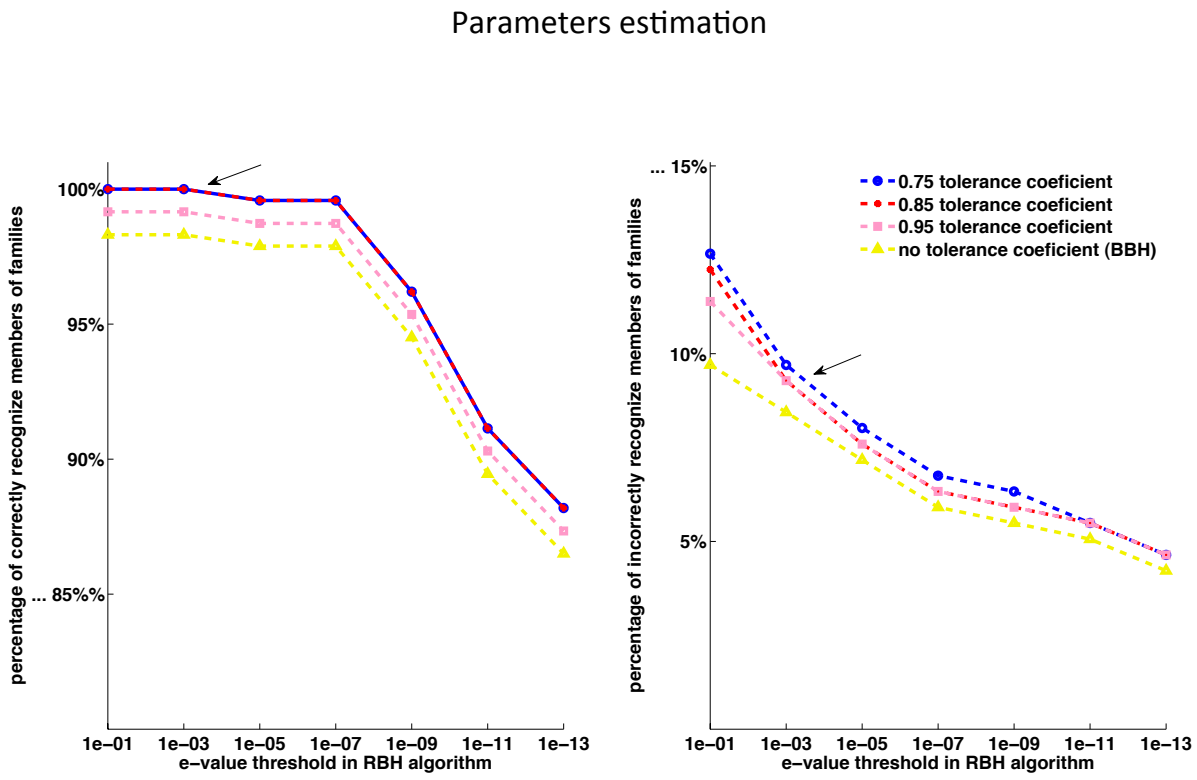


Figure 3.2. The estimation of the E-value threshold and the p-value tolerance coefficient parameters in the family-RDH algorithm. The percentage of the correctly recognized family members (using PHYLOMEDB families as a reference) for different values of the E-value threshold and the p-value tolerance parameters is shown on the left. The percentage of incorrectly recognized family members (using PHYLOMEDB families as a reference) for different values of e-value threshold and p-value tolerance parameters shown on the right. The arrow indicates the pair of parameters we decided to use for further analysis.

3.2.4.3 Testing for presence/absence of clade specific genes/gene families in Xenacoelomorph protein sets

The verified specific gene families (the families classified specific to Chordata, Deuterostomia, Bilateria (see Section 2.1, 2.2)) were further tested for the presence of the family members in

Xenacoelomorph protein sets (see Section 2.3) with the family-RBH algorithm. For each verified specific gene family we ran the family-RBH algorithm using Xenacoelomorph protein sets as the primary BLAST database and human genome as a secondary database. The algorithm verified if the top reverse BLAST hit is the member of the family; if yes, the gene was classified as present in the target Xenacoelomorpha protein set, otherwise as absent. If the sequence matched the human sequence with 100% identity we discarded it as a human conatamination. The results are represented quantitatively as a proportion of number of family members present in Xenacoelomorpha protein set to the number of families that were tested.

3.3 Results

3.3.1 Preliminary identification of Bilaterian, Deuterostome and Chordate specific gene families

We have identified 1,790 candidates for Bilateria specific gene families (acquired on the branch leading to Bilateria, found in Deuterostomia and in Protostomia but absent in all non bilaterian species), 42 candidates for Deuterostomia specific gene families (acquired on branch leading to Deuterostomia, found in Chordata and in Ambulacraria but absent in all non deuterostomes), 13,569 candidates for Chordata specific gene families (acquired on branch leading to Chordata, found in Vertebrata and non-vertebrate chordates but absent in non chordate deuterostomes and non deuterostomes) (see Methods 2.1).

3.3.2 Verifying absence of clade specific gene in outgroup taxa using NCBI database

To avoid including gene famiies which do have members in species not present in the PhylomeDB (as the composition of the proteomes included in PhylomeDB do not contain many sequenced species from Ambulacraria, Protostomia, Cnidarians and Ctenophores), the candidate gene families from PhylomeDB were confirmed for the absence of similar homologs by one directional remote BLAST search

in ncbi resources. The confirmation step significantly reduced the number of clade specific gene families to 220 in Bilateria, in 20 Deuterostomia, 257 in Chordata (see Methods 2.2).

3.3.3 Bilaterian specific gene families in Xenacoelomorpha

We have investigated the presence of 220 gene families specific to Bilateria in proteomes of *Meara stichopi* (ME)*, Nemertoderma westbladi* (NE)*, Pseudophanostoma variabilis* (PA) *Symsagittifera roscoffensis* (SY)*, Xenoturbella bocki* using family-RBH algorithm we developed specifically for that purpus (see Methods). We found 41% (90 gene families) Bilateria specific gene families in at least one Xenacoelomorpha. 25% of gene families were present in *Xenoturbella bocki* (XE). Other Xenacoelomorpha proteomes (*Meara stichopi (ME - 1), Nemertoderma westbladi (NE - 49), Pseudophanostoma variabilis (PA - 7), Symsagittifera roscoffensis (SY - 21)*) contained fewer Bilateria specific gene family. For comparison, we searched for Bilateria specific gene families in non-bilaterian *Nematostella vectensis* (NV), chordate *Ciona intestinalis* (CI), ambulacrarian *Strongylocentrotus purpuratus* (SP), *and protostome Apis mellifera* (AP). Out of 220 Bilateria specific gene families (acquired on the branch leading to Bilateria and present in Bilateria Last Comon Ancestor) we found more then 25% present in extant Bilaterians *Ciona intestinalis 25%*, *Strongylocentrotus purpuratus 47% Apis mellifera 35%*, (Figure 3.3), suggesting that they have been frequently lost since the PD divergence.

The result suggests that Xenacoelomorpha did not loose more of Bilateria specific gene families than other Bilaterians. The poor quality of Xenacoelomorpha proteomes likely decreased the recall performance, which is observed more strikingly in the Xenacoelomorpha species, which have low quality genomes (see Chapter 2). The presence of genes specific to Bilateria in Xenacoelomorpha at the same level as in other Bilaterians (deuterostomes and protostomes) may suggest that Xenacoelomorpha are in fact positioned within deuterostomes or protostomes. Ultimately these results cannot support nor can deny the position of the Xenacoelomorpha outside of bilaterians. Considering the low level of presence

of the Bilateria specific gene families in other bilaterians, it is not unexpected to observe the presence of 25% of bilaterian gene families in xenacoelomorphs.

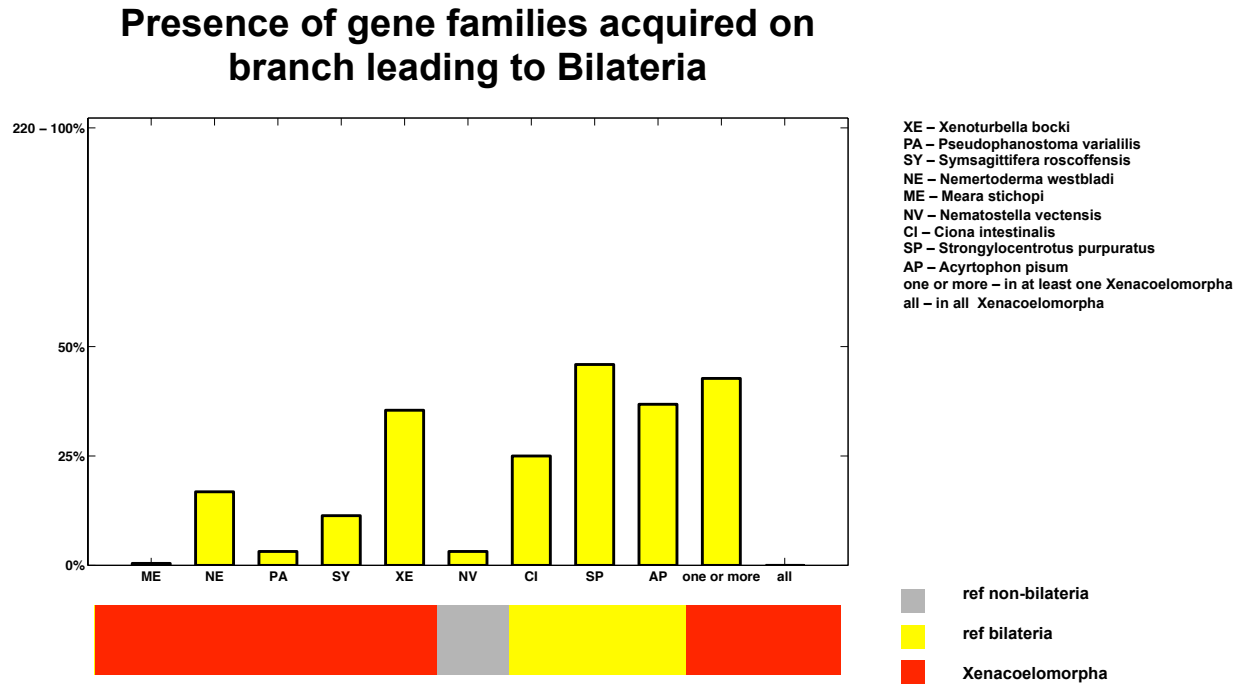


Figure 3.3. Bilaterian specific gene families are present in in Xenacoelomorpha proteomes. Histogram represents number of present gene families acquired on a branch leading to Bilateria in each of the contested species. Xenacoelomorpha marked by red bar *Meara stichopi* (ME)*, Nemertoderma westbladi* (NE)*, Pseudophanostoma variabilis* (PA) *Symsagittifera roscoffensis* (SY)*, Xenoturbella bocki* and non bilaterian *Nematostella vectensis* (NV) grey, chordate *Ciona intestinalis* (CI), ambulacrarian *Strongylocentrotus purpuratus* (SP), *and protostome Apis mellifera* (AP) yellow. 41% (90 gene families) were found in at least one Xenacoelomorpha. 25% of gene families are present in *Xenoturbella bocki* (XE) were reference Bilateria conatain: *Ciona intestinalis 25%*, *Strongylocentrotus purpuratus 47% Apis mellifera 35%.*

3.3.4 Deuterostome specific gene families in Xenacoelomorpha

After rigorously testing for the absence of homologs in out-group taxa we have found only 20 Deuterostomia specific gene families (acquired on the branch leading to Deuterostomia). While these are required to be present in both chordates and Ambulacraria, in a similar situation to that seen for the bilaterian specific characters it is by no means true that all are discoverable in the proteomes of all deuterostomes that we have investigated. Of these 20 deuterostome specific characters just 5 gene

families were present in the extant deuterostomes *Ciona intestinalis* and 10 in *Strongylocentrotus purpuratus* (see Figure 3.4) for example.

Seven of Deuterostomia specific gene families were found in more than one Xenacoelomorpha proteomes (see Figure 3.4). We found 6 of Deuterostomia specific gene families in *Xenoturbella bocki* (XE). Other Xenacoelomorpha protein sets (*Meara stichopi, Nemertoderma westbladi, Pseudophanostoma variabilis, Symsagittifera roscoffensis*) contain 0-2 Deuterostomia specific family members, which is likely due to poor data quality of the protein prediction datasets for this species (see Chapter 1). However, this indicates that deuterostome specific gene families are present in Xenacoelomorpha and could therefore suggest that Xenacoelomorpha belong to the deuterostomes.

Low total number of gene families acquired on the branch leading to deuterostomes, which is understandable considering how short a branch leading to Deuterostomes in comparison to branches leading to Chordata of Bilateria is, makes us less certain about the impact of our result on the phylogenetic position of Xenacoelomorpha. Even though the presence on gene families acquired on the branch leading to is suggestive, we could imagine a scenario where this gene families where acquired on the branch leading to Bilateria and then lost on the branch leading to Deuterostomia, if Xenacoelomorpha are basal bilaterians. Possibly, the presence of these genes in Xenacoelomorpha could be a result of the contamination, or the false positive result of the family-RBH algorithm.

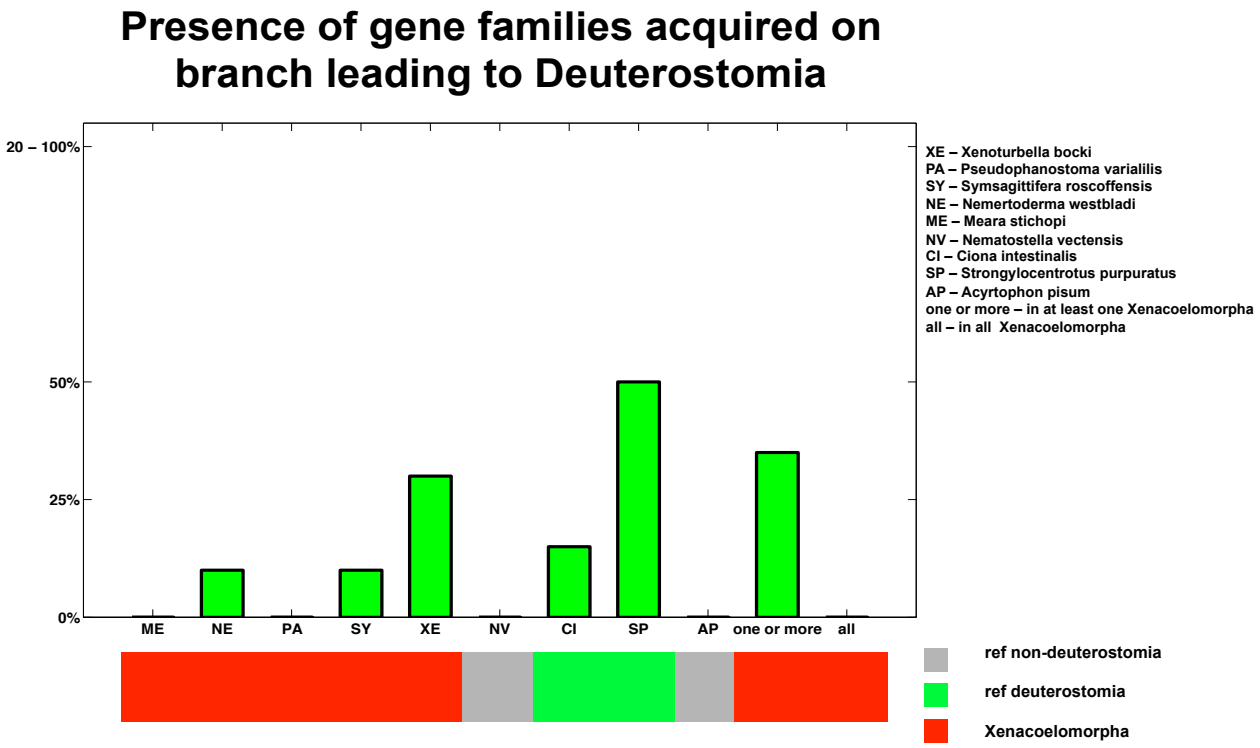


Figure 3.4. Deterostome specific gene families are present in Xenacoelomorpha proteomes. Histogram represents number of gene families present that were acquired on a branch leading to Deuterostomia in each of the contested species. Xenacoelomorpha marked by red bar *Meara stichopi* (ME), *Nemertoderma westbladi* (NE), *Pseudophanostoma variabilis* (PA) *Symsagittifera roscoffensis* (SY), *Xenoturbella bocki* and non bilaterian *Nematostella vectensis* (NV) grey, chordate *Ciona intestinalis* (CI), ambulacrarian *Strongylocentrotus purpuratus* (SP), *and protostome Apis mellifera* (AP) green. The presence of gene families present that were acquired on a branch leading to Deuterostomia in one or more Xenacoelomorpha on the right.

3.3.5 Apparent chordate specific gene families also present in Ambulacraria

Out of 257 verified Chordata specific gene families (present in Vertebrate and non-vertebrate chordates) 32% are found in an extant Urochordate (CI - *Ciona intestinalis*) (Figure 3.5). We were able to identify 20% of these gene families in at least one Xenacoelomorpha. Surprisingly, our BLAST analysis also showed that 20% of these gene families are identified in *Strongylocentrotus purpuratus*. This result must be an artefact of the computational pipeline.

The size of the BLAST database in the in depth similarity search is much bigger then the *Strongylocentrotus purpuratus* genome, therefore short fragmented sequences, that are insignificant hits

in large ncbi search, may still have passed the verification criteria and be significant when blasted against *Strongylocentrotus purpuratus* protein set. These sequence fragments may still be part of the gene that belongs to the same family. The poor quality of the *Strongylocentrotus purpuratus* genome and the few Ambulacraria species in PhylomeDB database and ncbi resources may not have been enough data required to remove gene families that are still present in Ambulacraria during the ncbi BLAST verification step. This highlights the limitations of one directional BLAST approach during the verification step. In this case we should consider this sequences as Deuterostome specific not Chordate specific. Thus, we can potentially consider this result as a positive evidence of deuterostome membership for Xenacoelomorpha. It would be worth testing if the same gene families found in *Strongylocentrotus purpuratus* are also found in at least one xenacoelomorph. Another possible explanation for the presence of supposedly Chordata specific gene families in *Strongylocentrotus purpuratus,* as well as in Xenacoelomorpha, could be contamination of the *Strongylocentrotus purpuratus* and Xenacoelomorpha genome data that we did not discover by the exact match to human sequences.

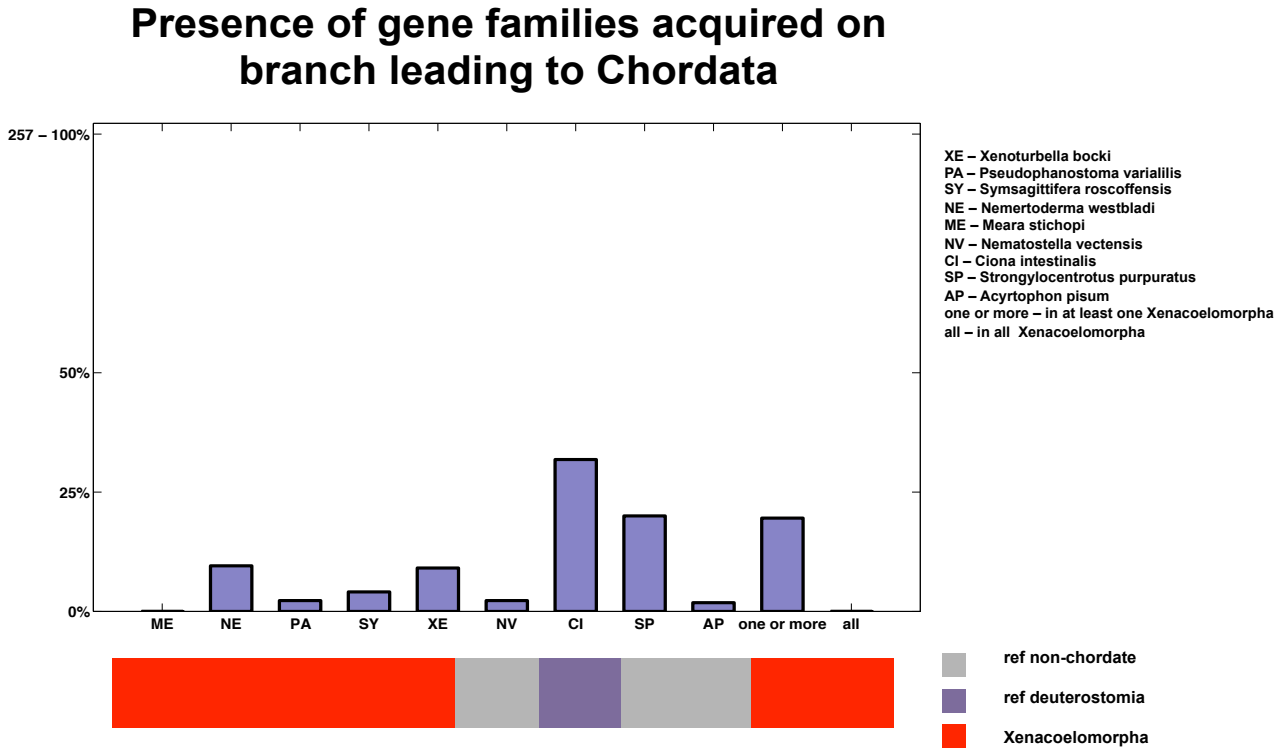


Figure 3.5. Apparent chordate specific gene families also present in Ambulacraria proteomes. Histogram represents number of present gene families acquired on a branch leading to Chordata in each of the contested species. Xenacoelomorpha marked by red bar *Meara stichopi* (ME), *Nemertoderma westbladi* (NE), *Pseudophanostoma variabilis* (PA) *Symsagittifera roscoffensis* (SY), *Xenoturbella bocki* and non bilaterian *Nematostella vectensis* (NV) grey, chordate *Ciona intestinalis* (CI), ambulacrarian *Strongylocentrotus purpuratus* (SP), *and protostome Apis mellifera* (AP) yellow. The presence of gene families present that were acquired on a branch leading to Chordata in one or more Xenacoelomorpha on the right.

3.3.6 An implication for the evolution of Xenacoelomorpha

We found Deuterostome specific genes present in Xenacoelomorpha and Bilateria specific genes present in Xenacoelomorpha. The presence of the members of gene families acquired on a branch leading to Deuterostomia (acquired on DLCA-BLCA branch) supports the phylogenetic position of Xenacoelomorpha within the deuterostomes, but not the protostomes. However, taking into consideration the small number of Deuterostomia specific families and small number of Ambulacraria genomes in PhylomeDB database and ncbi database, it is difficult to demonstrate more precise placement of Xenacoelomorpha using this method. A low proportion (between 0-50%) of clade specific gene families

was present in the extant animal proteomes (protein sets predicted based on genome and transcriptome assemblies). This is opposite to the ancestral gene families (see results 3.3.8) which frequently present in all axtant animals. This suggests that clade specific gene families are genes, which are frequently lost during evolution.

3.3.8 Correlation of the gene family loss with morphological complexity

It has been previously suggested that differential loss of ancestral gene families have played a role in the differentiation of animal phyla (Hughes 2004). We aim to estimate the level of ancestral gene families lost in Xenacoelomorpha and compare it to other extant organisms. Several attempts were made to estimate the rate of ancestral gene family loss in within the deuterostomes and within the protostomes in the past, reporting contradicting results (Friedman 2001; McLysaght 2002). However, it was suggested that the ancestral Bilaterian gene families are fairly conserved in the extant animals, and 70-90% of urbilaterina gene families are present in modern day genomes (Simakov et al. 2013). We have tested the set of gene families from PhylomeDB for the presence in the ancestor of Metazoa (present in both Bilateria and non-Bilateria Metazoa), ancestor of Bilateria (present in both Protostomia and Deuterostomia). Generated subsets of 13,556 ancestor Metazoa and 16,299 ancestor Bilateria gene families, were tested for the presence of the members in Xenacoelomorpha protein sets using family-RBH algorithm and reference protein sets of extant animals (see Figure 3.6,3.7).

We estimate that around 70-90% of ancestral Metazoa and Bilateria gene families are present in extant animals, which agrees with previous eatimation by Simankov. We inferred 13556 gene families in ancestor of Metazoa and 16299 gene families in the ancestor of Bilateria, suggesting that the number of gene families present in the common ancestor increases on the branch leading to Bilateria. Between 77-87% (10,438 – 11,794) ancestral Metazoa gene families are present in extant animal proteomes we investigated. More, but lower proportion, between 71%-83% (11,572 - 13,568) ancestral Bilateria gene families are present in extant animal proteomes.

The rate of ancestral gene family loss in high quality Xenacoelomorpha protein sets (see Chapter 1) does not differ significantly from other reference extant Bilaterians (see Chapter 1). We found over 90% of Metazoa and Bilateria ancestral gene families in at least one Xenacoelomorpha. Similar level of Metazoa and Bilateria ancestral gene families can be found in other extant organisms. We understand most of these gene families are widely present in animals and are probably cellular housekeeping genes. We found between 70-75% bilaterian, metazoan, deuterostomian ancestral gene families in at least one Xenacoelomorph. This corresponds to the results for other extant Bilateria and can be a result of lineage differentiation. This means that the apparent simplification of xenacoelomorphs was not a result of the loss gene families that were already present in Metazoa and Bilateria common ancestor, but likely the loss of other genetic characteristics like linage specific genes or paralog copies.

Because the level of gene family loss does not differ from other extant bilaterian organisms, there is no evidence suggesting that ancestral gene family loss was the reason for morphological simplification of Xenaceolomorpha. Xenaceolomorpha appear to posses the same level of Bilaterian gene families as other bilaterally symmetrical animals, as well as early branching non-bilateral Metazoa (*Nematostella vectensis*), which is in agreement with previous results (Putnam et al. 2007). Further investigation on a more detailed level is necessary to find more detailed molecular evidence about the gene loss on Xenacoelomorpha branch. We expect the reduction of paralog copies is a plausible reason for simplification of Xenacoelomorpha (which you are unable to detect using this method), rather than deletions of whole families that are crucial for proper functionalization of the organisms. Family RBH algorithm we used does not distinguish between paralogs and orthologs. The detection of orthology/paralogy relations within families need correct knowledge of species phylogeny, however the placement of Xenacoelomorpha on the tree of life is debated.

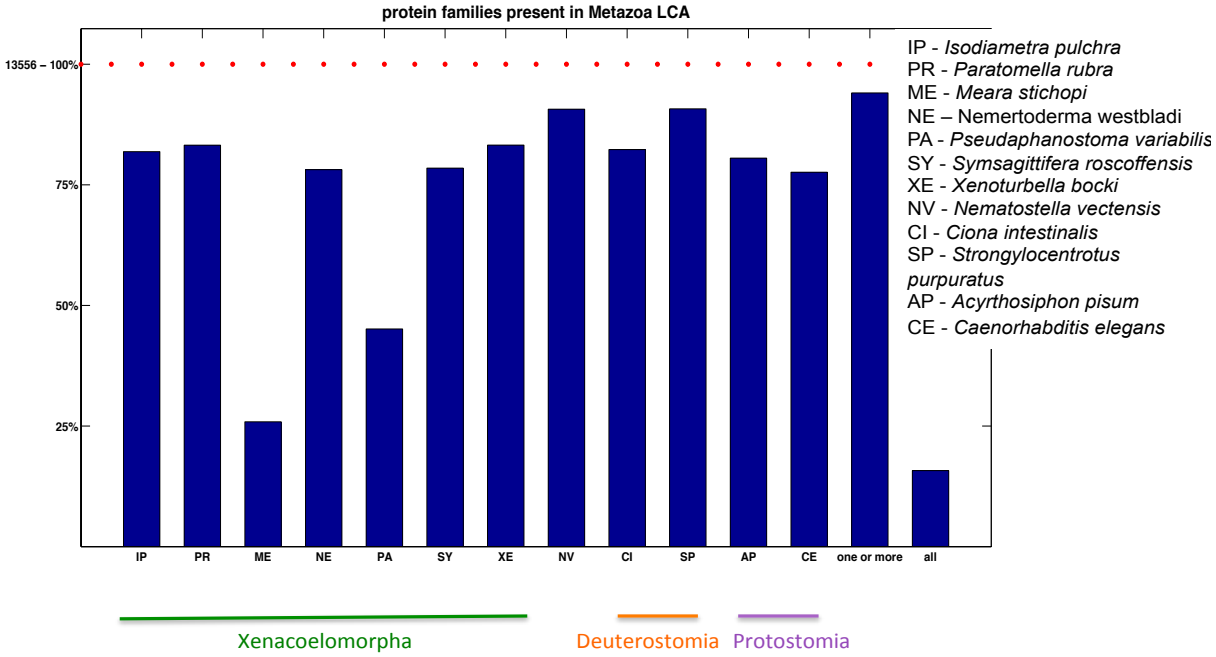


Figure 3.6. High proportion of Metazoa ancestral gene families is present in extant animals. Between 70-90% of ancestral Metazoa gene families are present in extant animal proteomes. We inferred 13556 gene families in ancestor of Metazoa. Little proportion of ancestral Bilateria gene families is present in proteomes of 25% *Meara stichopi* and 40% *Pseudaphanostoma variabilis*.

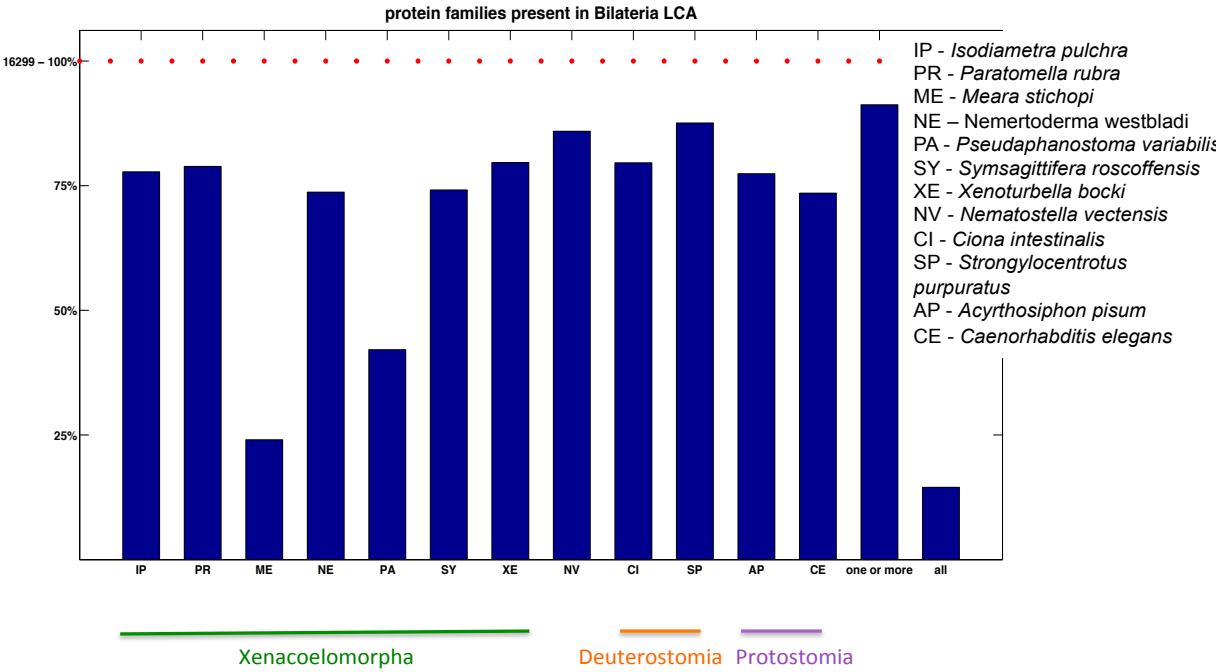


Figure 3.7. High proportion of Bilateria ancestral gene families is present in extant animals. Between 70-90% of ancestral Metazoa gene families are present in extant animal proteomes. We inferred 16299 gene families in ancestor of Bilateria. Little proportion of ancestral Bilateria gene families is present in proteomes of 23% *Meara stichopi* 39% and *Pseudaphanostoma variabilis*.

3.3.9 Conclusions

First and the biggest limitation of the approach we used, was the fact that all the gene families, which presence we have investigated in this Chapter, were human present gene families. This was the result of the way the phylomes were constructed in the database (based on human seed proteins). Such a priori condition allowed only specific set of questions to be asked. This condition did not allow us to check alternative scenarios, for the presence and absence of Ambulacraria specific gene families, which could be more informative for the position of Xenacoelomorpha.

Further limitations resulted from limited number of species in the PhylomeDB database, limited number of Ambulacraria species in the database and various data quality for each Ambulacraria and Xenacoelomorpha proteomes. Due to the possibility of incomplete sequence data, it was extremely difficult to conclude about the absence of the gene in the genome or protein set. Additionally, unknown levels of contamination, that was difficult to remove, in each dataset could influence the outcome of the analysis and made it hard to interpret.

One or bidirectional algorithms for orthology inference that are based on similarity searches do not perform well in cases of recent gene duplications or differential gene loss (hidden paralogy). We were not able to infer orthologous genes using PhylomeDBed resources and bidirectional similarity search approach. Mutually most similar sequences are often recognised as orthologous in the cases when multiple losses were involved during the evolution. Genes that have different length, evolve with different evolutionary rate, and undergo different frequency of losses and duplication events are treated with the same threshold, which decreases the performance of such algorithms.

In order to address these problems, in next Chapter, we have constructed a bigger dataset, which contained multiple species from Ambulacraria and other Metazoa. We used multidirectional approach that calculated evolutionary distance between all genes in multiple species, and accounts for differential gene loss, in order to identify sets of orthologs across Metazoa. There we used orthologous sets of genes to

infer the species phylogenetic tree (see Chapter 4 and 6) and we analyze the duplications, gains and losses within gene families (see Chapter 5).

Chapter 4

Impact of automated orthology group assignment on the reconstruction of lophotrochozoan phylogeny

4.1 Introduction

The phylogenetic position of Xenacoelomorpha is the subject of ongoing debate in the literature. Previous phylogenetic analysis of the orthologous genes inferred from the ESTs and genomic data using reciprocal approach (Philippe et al. 2011), orthoMCL clustering approach (Hejnol et al. 2009) and Hidden Markov Model gene profile approach (Cannon et al. 2016) are not with agreement with each other. We aim to improve on previous analysis by and selecting the best method of finding orthologs from animal genomes. We aimed to test the performance previously used methods (OrthoMCL, CEGMA) in constructing phylogenetic matrices from high throughput sequencing data together with multidirectional distance approach (OMA; Altenhof et a. 2014). As a test dataset we chose a subset of new generation sequencing data I used in Egger et al. 2015 paper. The work presented here will focus on the analysis of test dataset and Lophotrochozoa phylogeny.

Resolving the relationships of ancient lineages remains challenging for molecular phylogenetics. Interrelationships of the main clades of Lophotrochozoa, a large phylum containing several model organisms such as the cuttlefish *Sepia officinalis*, the snail *Lottia gigantea*, the earth worm *Lumbricus terrestris*, the freshwater leech *Helobdella robusta* and freshwater planarian *Schmidtea mediterranea* are debated in the literature (Hejnol et al. 2009; Smith et al. 2011; Kocot et al. 2011; Struck et al. 2011). Here, we attempted to improve the understanding of Lophotrochozoa phylogeny by collecting

Next Generation Sequencing data from various resources and establishing bioinformatics pipeline for constructing large protein alignment for phylogeny inference (supermatrix) (see Figure 4.1).

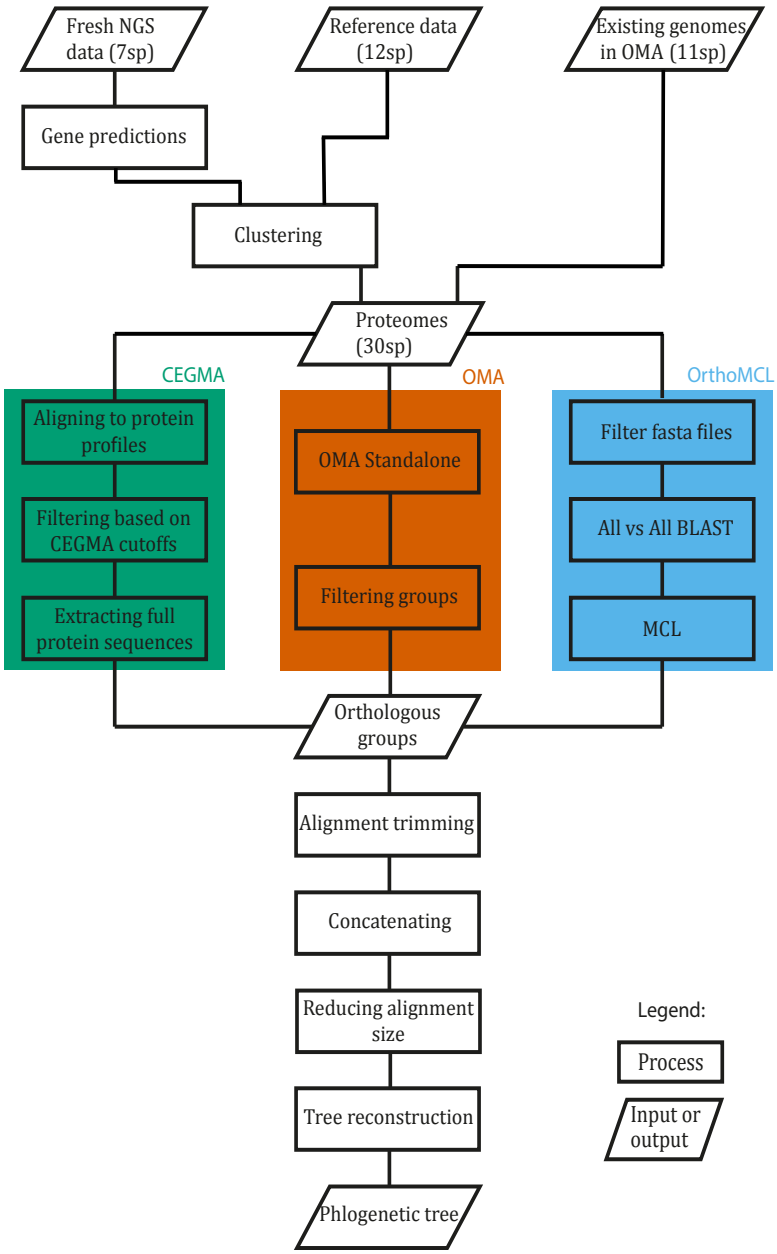


Figure 4.1. The flowchart of the OMA, CEGMA and OrthoMCL phylogenetic pipelines. The chart visualizes a set of procedures for all tree pipelines starting from raw genetic data and resulting in a phylogenetic tree. The data structures are marked by rhomboids, while the processes in which these data are used are marked by rectangles. Processes used to perform OMA, CEGMA and OrthoMCL orthology inference are marked in collared boxes.

Recently, the application of large phylogenetic matrices constructed from Next Generation Sequencing data helped to resolve parts of Lophotrochozoa phylogeny. Within Mollusca,

two recent large-scale analysis, support the monophyly of Conchifera (shell-bearing molluscs), and position Bivalia sister group to Gastropoda and basally positioned Cephalopoda (Smith et al. 2011; Kocot et al. 2011). Within Platyhelminthes (flatworms), two recent phylogenetic analysis revise the relationships between the clades divide the phylum into Catenulida and Rhabditophora, within Rhabditophora the earliest-emerging branch is Macrostomorpha, Lecithoepitheliata are sister group of Polycladida and Rhabdocoela as the most basally branching euneoophoran taxon with Proseriata sister group to Acentrosomata (Laumer et al. 2015a; Egger et al. 2015). Although the phylogeny of major clades is well resolved, the position of the basally branching clades, such as Nemertea (ribbon worms i.e. *Cerebratulus lacteus*) and Rotifera (commonly called wheel animals; i.e. model organism *Adineta ricciae*) is still debated (Hejnol et al. 2009; Smith et al. 2011; Kocot et al. 2011; Struck et al. 2011).

We attempted to resolve the phylogeny of some lophotrochozoan species, by collecting the data from Next Generation Sequencing, and gathering 30 most complete sets of proteins (proteomes) from 19 lophotrochozoans, 4 deuterostomes, 4 ecdysozoans and 3 non-bilaterian species for the analysis. We used an RNA-seq approach to generate new transcriptomes for 7 lophotrochozoan species (also used in Egger et al. 2015; species names marked with blue on Figure 4.7,4.8,4.9,4.10,4.11,4.12). Additionally, we included 12 selected species available in the NCBI Reference Sequence Database (http://www.ncbi.nlm.nih.gov/refseq/; species names marked with grey on Figure 4.7,4.8,4.9,4.10,4.11,4.12) and 11 genomes available on the OMA export page (http://cbrg-oma-test.ethz.ch/oma/export/; species names marked with black on Figure 4.7,4.8,4.9,4.10,4.11,4.12). For our new data, we assembled the transcriptomes, generated protein predictions and removed redundant sequences from the dataset. From this data we aimed to construct the longest and the most complete large phylogenomic matrices (supermatrices) by comparing 3 phylogenetic pipelines for supermatrices construction. It was previously shown that such supermatrices, if complete, are informative in resolving the deepest nodes in the tree of life (von Reumont et al. 2012; Fernández et al. 2014; Laumer et al. 2015a). However, the proper construction of such supermatrices remains a challenge for large-scale phylogenetic analysis (Dunn et al. 2008; Pick et al. 2010).

Lophotrochozoa					
Platyhelmintha	Gastrotricha	Mollusca	Nemertea	Analida	Rotifera
Schmidtea mediterranea	*Mesodasys laticaudatus*	*Biomphalaria glabrata*	*Cerebratulus sp.*	*Lumbricus rubellus*	*Brachionus plicatilis*
Monocelis sp.		*Lymnaea stagnalis*		*Helobdella robusta*	*Adineta ricciae*
Microdalyellia schmidt		*Lottia gigantea*		*Capitella teleta*	
Echinoplana celerrima		*Sepia officinalis*			
Macrostomum lignano		*Chaetopleura apiculata*			
Catenula lemnae					
Ecdysozoa	Deuterostomia	Non-bilateria			
Pristionchus pacificus	*Saccoglossus kowalevskii*	*Hydra magnipapillata*			
Caenorhabditis elegans	*Strongylocentrotus purpuratus*	*Trichoplax adhaerens*			
Acyrthosiphon pisum	*Ciona intestinalis*	*Amphimedon queenslandica*			
Drosophila melanogaster	*Homo sapiens*				

Table 4.1 Next Generation Sequencing data for 30 species used in the analysis. The dataset includes 30 most complete sets of proteins (proteomes) from 19 lophotrochozoans, 4 deuterostomes, 4 ecdysozoans and 3 non-bilaterians.

Supermatrices are the concatenated constructs of the sequence alignments, which are built from the groups of orthologous genes. The use of large (multigene) dataset drastically reduces the random (or sampling) error in phylogenetic reconstruction. However, it should be noted that poor taxon sampling, missing data (gene coverage and incomplete sequences) erode statistical power and sometimes enhance tree reconstruction artefacts (often species with lots of missing data are artificially grouped together). Phylogenomic datasets, especially when based on expressed sequence tag (EST) data (Lemmon et al. 2009; Philippe et al. 2004 Philippe et al. 2011), are frequently characterised by incomplete gene coverage for some taxa. Moreover, automated methods for orthology and paralogy prediction have to be used when the phylogenetic analysis is performed on the genome scale, but these methods have limited performance and manual curating is necessary (Philippe et al. 2011). For example, automated methods such as reciprocal best-hit or clustering methods are particularly susceptible to cases of hidden paralogy. These cases are often caused by differential gene loss or missing data in genomic sequences, because of that the most similar sequences may not necessarily be orthologous. It is vital that our orthology inference approach assembles datasets of genes, which

are all orthologous to one another, because the inclusion of paralogs in the construction of phylogenetic matrices can have a detrimental effect on the phylogeny reconstruction. When constructing a phylogenetic tree from paralogous sequences, the tree would represent the evolutionary distance between a duplication event, rather than speciation. Including paralogs in the dataset for the animal tree reconstruction analyses can lead to the incorrect inference of the relationships between species (Doyle et al. 1992) (see Figure 4.2).

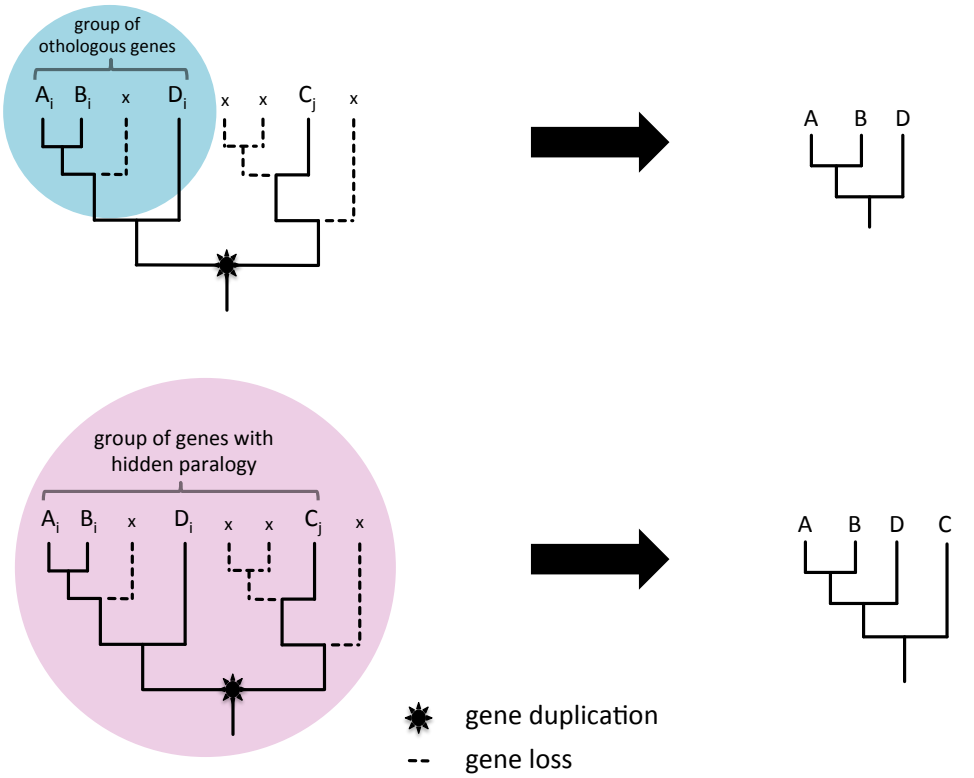


Figure 4.2. Phylogenetic tree inference based on a groups of orthologous (blue) and paralogous sequences (red). Phylogenetic tree constructed based on groups of orthologous genes (Ai,Bi,Di) represents the relation between species A,B and D (blue). The inclusion of difficult to detect by automated methods paralogs (Cj) in the group of genes for phylogenetic inference results in phylogenetic tree that does not represent the relation between species A,B,C and D (red).

To construct the best quality supermatrix, we compared the performance of 3 most popular automated methods for orthology inference (OMA standalone (Orthologous Matrix), the algorithm which uses evolutionary distances instead of scores, considers distance inference uncertainty, includes many-to-many orthologous relations and accounts for differential gene losses

(Altenhoff et al. 2015), OrthoMCL (Ortholog groups Markov Clustering) the algorithm for grouping proteins into ortholog groups based on their sequence similarity using Markov CLustering (Li et al. 2003) and CEGMA (Core Eukaryotic Genes Mapping Approach) the algorithm which uses Hidden Markov probabilistic gene models for finding orthologous genes in the proteomes (Parra et al. 2007). Due to the fact that missing data enhance tree reconstruction artefacts and long alignment helps to avoid stochastic error, we first tested amount of missing data produced by each method by investigating gene coverage and length and amino acid density of produced supermatrices. Next, because an inclusion of paralogs can lead to the incorrect tree inference, we tested how many genes reconstruct the monophyly of current accepted taxonomic clades. We chose the best performing method for finding orthologs and used it to reconstruct Lophotrochozoa phylogeny with both Maximum Likelihood and Bayesian approach. We compared obtained Maximum Likelihood and Bayesian phylogenetic trees with current state of knowledge about Lophotrochozoa evolution and with trees obtained with other two automated methods for orthology inference.

We showed that orthology inference, using the OMA pipeline, results in many more total number of orthology groups with at least 50% gene occupancy, than other two automated methods (CEGMA and OrthoMCL pipeline) for orthology inference, as well as more orthology groups than previous high throughput analysis based on Next Generation Sequencing data (Smith et al. 2011; Hejnol et al. 2009; Fernández et al. 2014). Furthermore, OMA pipeline generates more orthology groups between 50% and 75% gene occupancy than other two tested automated pipelines, and more groups between 75% an 100% gene occupancy. We showed that, OMA pipeline generates supermatrices with higher amino acid density, than the other two pipelines. Moreover, we show that significantly more gene trees, calculated from the OMA orthology groups, recover the monophyly of Lophotrochozoa clade and other currently accepted taxonomic animal groups more consistently than gene trees calculated from the CEGMA and OrthoMCL orthology groups. This result suggests a higher degree of paralogy in CEGMA and OrthoMCL datasets, as orthologous genes recover the monophyly of animal clades more often than groups containing paralogous genes.

Finally, we inferred the phylogeny of a subsample of the lophotrochozoan species, based on the supermatrices generated with OMA standalone (Orthologous MAtrix) pipeline (see Methods). We calculated inferred phylogeny using a Bayesian approach with CAT+GTR+Γ model in PhyloBayes (Lartillot et al. 2009) and Maximum Likelihood approach using the GAMMA GTR model implemented in RaXML. We show that the OMA standalone methods result in better-supported animal tree using both RaXML and PhyloBayes (Stamatakis 2014; Lartillot et al. 2013) and that is more consistent with current literature. We highlight the differences between the phylogenetic trees obtained with alternative pipelines and OMA standalone pipeline.

4.2 Materials and methods

4.2.1 Transcriptome assembly and peptide prediction

After quality assessment with FastQC it was determined using PRINSEQ lite (Schmieder and Edwards 2011) that the first 12 nucleotides needed to be trimmed off the 100bp (readshttp://www.bioinformatics.babraham.ac.uk/projects/fastqc/). Assembly of the trimmed paired reads was done using Trinity v20130225 (Haas et al. 2013) using the flag '--min_kmer_cov 2' in addition to default parameters. To test for the presence of cross contamination between libraries run on the same flow cell, we used the bowtie software and a custom script to identify any assembled transcript with fewer than four read matches which were discarded. In addition, we discarded all transcripts in which the number of reads from the intended species matching the transcript was not at least 5 times greater than the number of matches to the transcript from reads from any of the other potentially contaminating species. For peptide predictions, the Trinity script 'transcripts_to_best_scoring_ORFs.pl' was run on the nucleotide assembly, keeping all ORFs >100aa. For all peptide datasets cd-hit was used to reduce redundancy by clustering sequences with a global sequence identity of >95%.

4.2.2 Sequence Processing

Transcriptome reads from the following 7 previously unsequenced species, *Mesodasys laticaudatus* (Gastrotricha), *Catenulida sp.*, *Macrostomum lignano*, *Echinoplana celerrima*, *Microdalyellia schmidti*, *Monocelis sp.* (Platyhelminthes) and *Cerebratulus sp.* (Nemertea) were assembled as described (Grabherr et al. 2011). 12 sets of genomic and transcriptomic protein predictions from *Saccoglossus kowalevskii*, *Brachionus plicatilis*, *Adineta ricciae*, *Schmidtea mediterranea*, *Lumbricus rubellus*, *Chaetopleura apiculata*, *Sepia officinalis*, *Mytilus californianus*, *Biomphalaria glabrata*, *Lymnaea stagnalis, Hydra magnipapillata* and *Amphimedon queenslandica*, were downloaded from the NCBI refseq respository (ftp://ftp.ncbi.nlm.nih.gov/refseq/). Redundant sequences with higher than 97% identity were removed by clustering with CD-HIT (Limin et al. 2012). Additionally, 11 precomputed proteomes for *Homo sapiens*, *Strongylocentrotus purpuratus*, *Ciona intestinalis*, *Trichoplax adhaerens*, *Pristionchus pacificus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Acyrthosiphon pisum*, *Capitella sp.*, *Helobdella robusta* and *Lottia gigantea* were downloaded from the OMA database website. The combined set of 30 non-redundant protein sets contained 4 deuterostomes, 4 ecdysozoans, 19 lophotrochozoans and 3 non-bilaterian proteomes.

4.2.3 CEGMA pipeline

458 hidden Markov model protein profiles were downloaded from CEGMA (Parra et al. 2007). We aligned each component protein sequence to each core protein profile using hmmsearch (HMMER (http://hmmer.janelia.org)) and recorded the score. A protein sequence from a given species was retained if the alignment score matched the ortholog specific alignment and completeness cutoffs provided by CEGMA. If more than one predicted protein had a score above the relevant cutoff, the protein with the highest scoring alignment to the profile was selected as the 'true' ortholog. For each CEGMA profile, the single selected ortholog from all species (if any) were grouped as CEGMA core ortholog groups (see Figure 4.2).

4.2.4 OMA pipeline

The OMA standalone version 0.99w was downloaded from http://omabrowser.org together with 9 complete proteomes and precomputed pairwise alignments between every pair of proteins of all 9 proteomes. Our additional proteins sets were added and OMA orthologous groups were calculated using default parameters on a UCL computer science cluster (http://bioinf.cs.ucl.ac.uk/). OMA orthologous groups in which more than 50% of species were represented were selected for further analysis (see Figure 4.2).

4.2.5 OrthoMCL pipeline

30 non-redundant protein sets were further filtered with orthomclFilterFasta. NCBI blastp was run locally on preformatted (formatDB) sets of all proteins (Madden et al. 2013). The percent match length was computed and matches with E-Value < 1e-5 were kept. The potential in-paralog, ortholog and co-ortholog pairs were identified using the Orthomcl Pairs program and parsed for clustering. The mcl program was run, to cluster similar proteins into orthology groups with the inflation index 2.2 (Li et al. 2003). A FASTA file for each orthology group was created using get_seq_from_genomes_to_o_groups.pl.

4.2.7 Protein sequence alignments and Phylogenetic Analyses

Protein sequences from each orthology group containing sequences from at least 15 species were aligned using MUSCLE (Edgar et al. 2004), with default settings. Unreliable portions of the alignment were removed from the alignments using trimAl (Capella-Gutierrez et al. 2009), with default settings. The final alignment was created by concatenation of all alignments from 2,162 OMA orthology groups with 15 or more genes (>=50% complete, similarly we concatenated 438 CEGMA core orthology groups and 484 OrthoMCL groups with 15 or more members). Missing sequences were represented by gaps. The full alignment was finally reduced to sites with more then 60% occupancy, resulting in 386,499 aligned amino acid positions included from the OMA pipeline, 127,340 and 48,286 positions with CEGMA and OrthoMCL pipeline respectively.

Using these alignments, an ML analysis was conducted for both the OMA, CEGMA and OrthoMCL alignments using RAxML 8.0.14 (Stamatakis et al. 2014). Best-scoring ML trees were inferred using the protein GAMMA + GTR model from 100 replicate parsimony starting trees. The trees were inferred for 1000 bootstrap samples were and annotated at the best tree. Bayesian inference was conducted with PhyloBayes (PhyloBayes version1.5a in open mpi version 1.8.1 environment on a UCL Computer Science Cluster) using the CAT GTR model, in parallel using 32 CPU cores per chain (Lartillot et al. 2013). Two independent MCMC chains of 1,000 generations each were run on each alignment. The first 100 trees (10%) were discarded as burn-in for each MCMC run prior to convergence (i.e., when maximum discrepancies across chains <0.3).

4.3 Results

4.3.1 Comparison of orthology groups' sizes inferred with OMA, CEGMA and OrthoMCL pipelines

We generated 98,222 orthology groups using the OMA standalone pipeline. For further analyses we used a subset of 2,162 orthology groups that are present in at least 15 taxa, which corresponds to a 50% gene occupancy threshold. Previous large scale phylogenetic analysis within the Metazoa, which have opted to compute a supermatrix in order to resolve species phylogeny, managed to generate far fewer orthology groups with same gene occupancy threshold (53 orthology groups, among 94 species had 50% gene occupancy (Hejnol et al. 2009), 301 orthology groups, among 40 species had 50% gene occupancy (Smith et al. 2011)). Only the attempts that included fewer taxa and had a smaller phylogenetic spectrum have managed to obtain similar numbers of orthology groups with high gene occupancy (2,637 orthology groups, among 18 species of spiders had 62.47% gene occupancy (Fernández et al. 2014); 2,779 orthology groups among 18 species of Eutrochozoa (Nemertea, Mollusca and Annelida) had 78% gene occupancy (Andrade et al. 2014)).

In contrast, using the same starting set of 30 proteomes we used the CEGMA pipeline and OrthoMCL pipelines (see Methods), which are commonly used tools for discovering orthologs for phylogenetic analysis (Brejová 2009; Xu et al. 2014; Roy et al. 2014). We obtained fewer orthology groups with both CEGMA (458) and OrthoMCL (484) with minimum 50% gene occupancy (see Figure 4.3). For medium sized groups (groups containing between 15 and 25 genes), OMA yielded 2,111 groups, whilst CEGMA produced 271 and OrthoMCL 471. However, for larger sized groups (groups containing between 26 and 30 genes), OMA found only 51 groups compared to 171 found by CEGMA and 13 found by OrthoMCL. CEGMA approach uses HMM profiles of core 458 core genes that are present in 7 model organisms (Parra et al. 2007), thus 458 is the maximum number of orthologous groups that we could retrieve. The process of classification of best HMM hits to CEGMA groups relies

on the worst score in one of the 7 model organisms and can be susceptible to false positive orthology inference, and can result in the inclusion of paralogs (Berglund et al 2008). In contrast, in both OMA and OrthoMCL pipelines there is no limit of orthology groups that can be found, the algorithms analyse the relation between every protein in a genome, by relying on the all-to-all similarity search. In OMA standalone, the evolutionary distance is calculated during all-to-all similarity search, and proteins are classified into orthology groups, if they have a verified orthology relation to all the other members of the group (be the closest evolutionary distance protein among once species genome and has no witness of non-orthology). Similarly, in OrthoMCL significant reciprocal best hits, from all-to-all similarity search, are divided into groups using MCL (Markov Clustering algorithm Van Dogen 2000) clustering algorithm, which joins the most similar sequences into orthology groups.

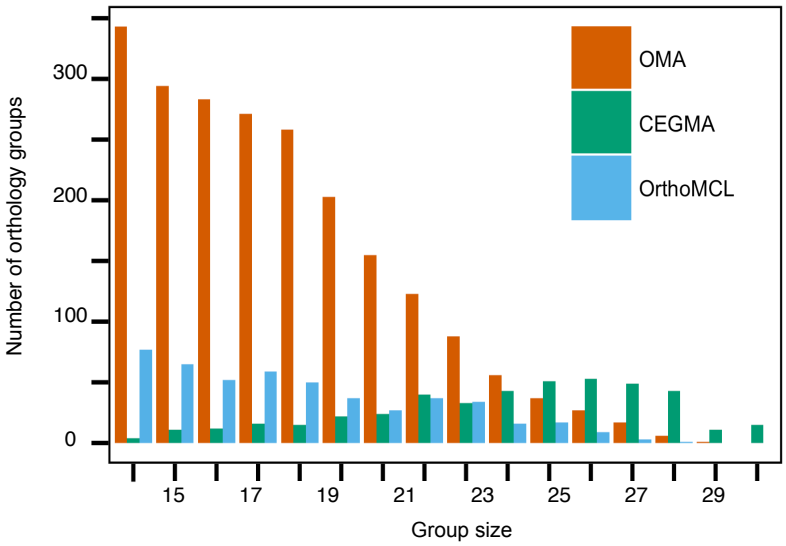


Figure 4.3 Distribution of group sizes. Using the OMA pipeline we inferred 2162 orthology groups with 15 or more genes, whereas with the CEGMA pipeline we inferred 442 orthology groups, and 484 orthology groups with the OrthoMCL pipeline. More CEGMA groups have 25 or more members in comparison to both OMA and OrthoMCL Most OMA and OrthoMCL orthology groups contain few genes (fewer than 20).

4.3.2 Supermatrix density comparison

The supermatrix inferred by the OMA pipeline (see Figure 4.2) was found to be more complete, than that of CEGMA and OrthoMCL (see Figure 4.4). The OMA pipeline produces more sites then CEGMA and OrthoMCL in the alignment that have between 40% to 80% occupied amino acid

positions. Only CEGMA produces more alignment positions with 80-100% occupancy, however, only 19,240 positions reach this threshold. We obtained a 136,499 position superalignment with less then 30% gaps on each of the positions, which was significantly more than with any other pipeline we used and in any of the previous attempts that didn't use OMA (Hejnol et al. 2009; Smith et al. 2011; Fernández et al. 2014; Andrea et al. 2014).

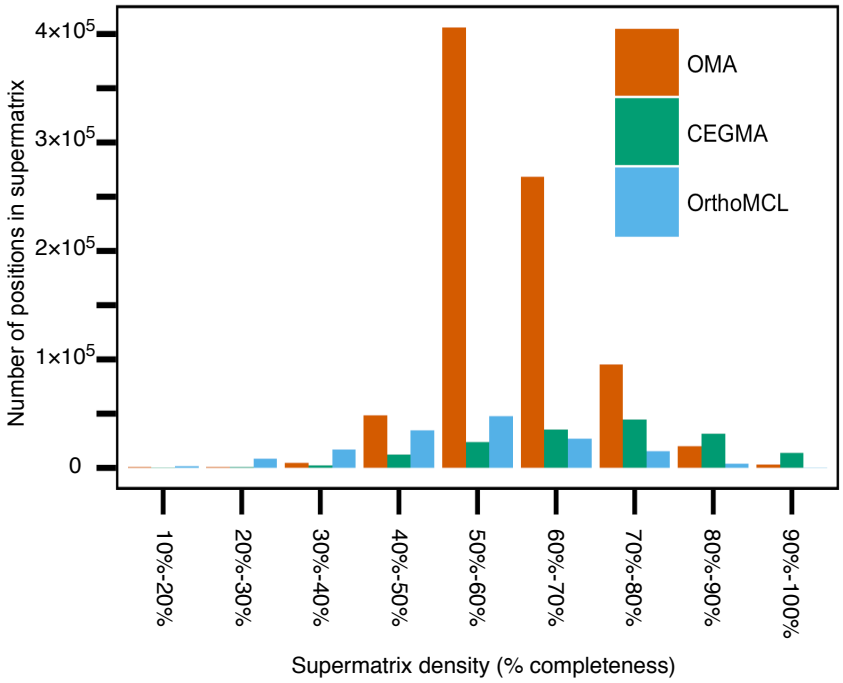


Figure 4.4. Histogram represents the number of amino acid positions with different supermatrix density. A supermatrix produced with the OMA pipeline has more positions with the density between 50-80%, while more positions between 80%-100% density are produced with CEGMA. OrthoMCL produces the supermatrix with the fewest positions between 60%-100%, while the most positions between 10%-30%.

4.3.3 Consistency with current taxonomy

In order to assess the quality of the orthology groups, we tested how consistent the gene trees produced by the individual orthology groups were with accepted major animal clades. To allow a direct comparison between the three pipelines, we selected, at random, an equal in size sample of OMA, CEGMA and OrthoMCL orthology groups of each size (between 15 and 30 species), and constructed gene trees using PhyML (Guindon et al. 2010). We then calculated the proportion of correctly placed taxa within 11 accepted major animal clades for each pipeline (see Figure 4.5). The orthologous genes identified by the OMA pipeline tend to support the established

clades more consistently than either CEGMA or OrthoMCL (see Figure 4.5). More gene trees based on OMA groups recover consensual animal clades as monophyletic, than trees based on either CEGMA or OrthoMCL groups. Overall, OMA gene trees recovered consensual animal clades as 62.09% monophyletic, whilst CEGMA recovered 56.37% and OrthoMCL recovered 52.42%. A Mann-Whitney U test was used to test if OMA orthology groups tends to recover the consensual animal clades as monophyletic more frequently then both CEGMA and OrthoMCL piplines. The difference in reconciliation of the Lophotrochozoa clade, where the clade has as much as 19 species in our dataset (other tested clades had form 6 to 2 species), was the most significant (α=0.05 between OMA and CEGMA, and 0.005 between OrthoMCL versus both OMA and CEGMA with Mann-Whitney U test). Similar results were obtained were all OMA, CEGMA and OrthoMCL groups with over 50% gene occupancy were compared (as opposed to a random sample equal in size, data not shown). The higher proportion of correctly recovered clades illustrates better performance in terms of the ability to recognise orthologs, as orthologous genes should recover animal phylogeny more frequently. This result suggests that the OMA orthology groups are more taxonomically informative than groups identified with CEGMA and OrthoMCL (except for Deuterostomia, Annelida, Arthropoda and Ambulacraria, but this clades had only 4 to 2 species in our dataset).

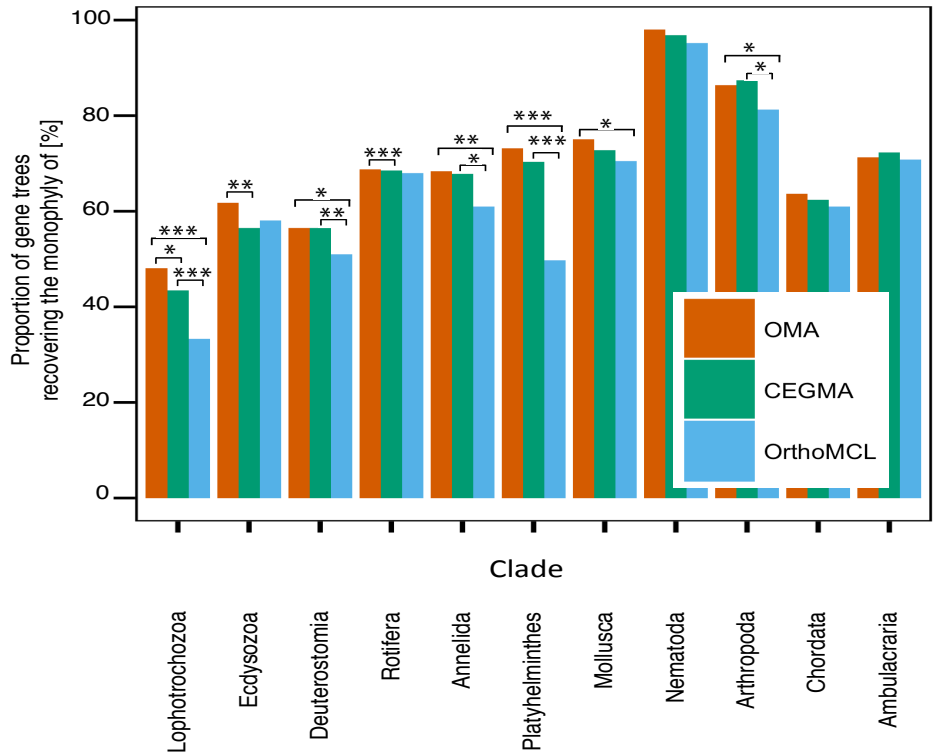


Figure 4.5. OMA orthology groups tend to recover the monophyly of Lophotrochozoa more frequently. Proportion of gene trees, calculated with PhyML from 220 randomly sampled OMA, CEGMA and OrthoMCL groups, with the same proportion for each group size, which recover consensual animal clades as monophyletic (were significance at α=0.05; 0.005; and 0.0005 is indicated with *, ** and *** respectively).

There could be several reasons why gene trees calculated with OMA orthology groups tend to recover the monophyly of animal clades more often. To better explain that, we have illustrated two examples of corresponding OMA and CEGMA orthology groups, where Lophotrochozoa clade are monophyletic at the phylogenetic tree calculated with OMA but not with CEGMA orthology group. First, lack of information on the intermediate steps in gene evolutionary history caused by missing sequences from the orthology group result in difficulties in gene tree inference. In example, in the OMA group 295, where gene sequence from *Catenulida lemnea* and *Hydra magnipapilata* are present in the orthology group, Lophotrochozoa are monophyletic and gene tree reflects the evolution of Metazoa (see Figure 4.6A; additional sequences indicated in green, different sequences indicated in blue, Lophotrochozoa indicated in purple). In the corresponding CEGMA orthology group 18, where gene sequence from *Catenulida lemnea* and *Hydra magnipapilata* are missing from the orthology group, Lophotrochozoa are paraphyletic (see Figure 4.6B). Moreover, different and slower evolving human sequence is present in the OMA

group 295 (indicated by the branch length). Second, paralogous sequences can be added to the orthology group instead of orthologous sequence or if the orthologous sequence was lost. Such case can be observed in OMA group 1272 and corresponding CEGMA group 1350 (see Figure 4.6 C and D). In CEGMA group 1350 paralogous sequences from *Catenulida lemnea* and *Mesodaysys laticaudatus* are added to the orthology group, instead of the orthologous once chosen by OMA standalone, which results in paraphyletic Lophotrochozoa.

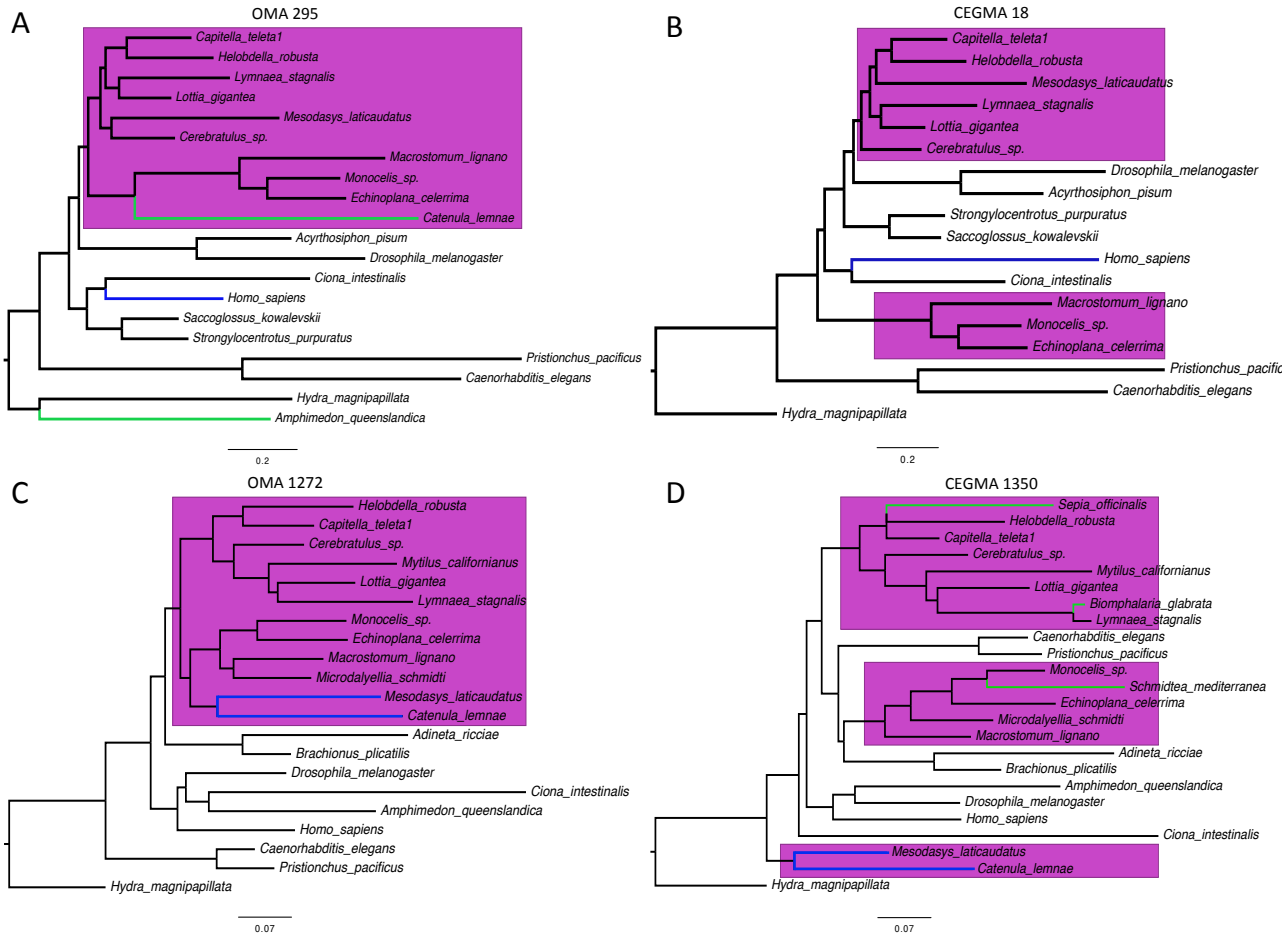


Figure 4.6. The example of gene trees, in which Lophotrochozoa are monophyletic on a gene tree calculated based on OMA orthology group, but not in the corresponding CEGMA orthology group. A) OMA group 295, where gene sequence from *Catenulida lemnea* and *Hydra magnipapilata* are present in the orthology group, Lophotrochozoa are monophyletic and gene tree reflects the evolution of Metazoa. B) CEGMA orthology group 18, where gene sequence from *Catenulida lemnea* and *Hydra magnipapilata* are missing from the orthology group, Lophotrochozoa are paraphyletic. C) OMA group 1272gene tree reflects the evolution of Metazoa D) In CEGMA group 1350 paralogous sequences from *Catenulida lemnea* and Mesodaysys laticaudatus are added to the orthology group (additional sequences indicated in green, different sequences indicated in blue, Lophotrochozoa indicated in purple).

4.3.4 Lophotrochozoa phylogeny inference

Next Generation Sequencing data largely improve the understanding of Lophotrochozoa evolution (Hejnol et al. 2009; Struck et al. 2011; Egger et al. 2015; Laumer et al. 2015a; Smith et al. 2011), however parts of Lophotrohozoa phylogeny are still debated. New evidence seems to resolve the systematics of Mollusca, two recent large-scale analysis, support Bivalia sister group to Gastropoda and basally positioned Cephalopoda (Smith et al. 2011; Kocot et al. 2011). Two recent phylogenetic analysis seems to be in agreement about the evolution of Platyhelminthes. Authors divide the phylum into Catenulida and Rhabditophora, within Rhabditophora the earliest-emerging branch is Macrostomorpha, Lecithoepitheliata are sister group of Polycladida and Rhabdocoela as the most basally branching euneoophoran taxon with Proseriata sister group to Acentrosomata (Laumer et al. 2015a; Egger et al. 2015). Although the phylogeny of major clades is well resolved, the position of the basally branching clades, such as Nemertea (ribbon worms i.e. *Cerebratulus lacteus*) and Rotifera (commonly called wheel animals; i.e. model organism *Adineta ricciae*) is still debated (Hejnol et al. 2009, Smith et al. 2011, Kocot et al. 2011; Struck et al. 2011). To approach this problem, we collecting the data from Next Generation Sequencing, and gathering 30 most complete sets of proteins (proteomes) from 19 lophotrochozoans and 11 other metazoans to gather (see Figure 4.7; new data species names marked with blue, OMA export page data species names indicated in black, refseq data indicated in grey).

Based on the results of the gene occupancy and the monophyly test of orthology groups, as well as the amino acid supermatrix (the concatenation of the alignments of multiple orthology groups) density results, we conclude that OMA standalone produces the supermatrix of the best quality. Therefore, we chose the supermatrix produced using OMA standalone pipeline, as the most suitable for phylogenetic analysis of Lophotrochozoa. We use this supermatrix to perform the phylogenetic analysis of Lophotrochozoa using both Bayesian and Maximum Likelihood method. We compare the results with previous phylogenetic analysis and current knowledge about Lophotrochozoa evolution. Next, compare the results with the phylogenies obtained based on CEGMA and OrthoMCL

supermatrices, using the same models of molecular evolution with both Bayesian and Maximum Likelihood methods. We discuss the differences between these phylogenies and assess which one is more consistent with recent literature.

We inferred phylogenetic tree of Lophotrochozoa using site heterogeneous CAT+GTR+ Γ model of molecular evolution (see Chapter 6) with the Bayesian method, based on the OMA orthologous gene set. The obtained phylogeny has high support (all pp=1 except the nodes involving the branching of Rotifera to Platyhelminthes and Cephalopoda to Bivalvia and Gastropoda) (see Figure 4.7). Our OMA Bayesian tree supports previous phylogenetic studies and divides of Lophotrochozoa into Rotifera (commonly called wheel animals a phylum of microscopic and near-microscopic pseudocoelomate animals) and other Lophotrochozoa (Struck et al. 2008; Philippe et al. 2011; Egger et al. 2015). Sister clade to Rotifera (cyan; *Adineta ricciae*, *Brachionus plicatilis*) consists of two monophyletic groups. First group (green) consists of Gastrotricha (worm like pseudocoelomate animals, commonly called hairybacks) being grouped together with Platyhelminthes (flatworms, acoelomate animals without circulatory and respiratory organs). Second group consists of Annelida (triploblastic coelomate segmented worms with circulatory system; brawn) being grouped together with Mollusca (coelomate unsegmented animals with circulatory system; orange) and Nemertea (acoelomate ribbon worms with circulatory system, magenta).

Our Bayesian analysis of OMA dataset places Nemertea as a sister group to Molusca, and supports previous findings by Hejnol et al. (2009), Struck et al. (2008) and Kocot et al. (2011). Moreover, our result supports phylogenetic analysis of Struck et al. (2011) with low degree of missing data and the jacknife analysis by Egger et al (2015). However, the placement of Nemertea is still debated in the literature and other evidence provided by authors (Laumer et al. 2015a; Egger et al. 2015; Struck et al. 2014) place Nemertea as a sister group to Annelida is a notoriously problematic taxon for phylogenetic classification (Struck et al. 2008).

Both the maximum likelihood and Bayesian trees show support for the position of Nemertea as a sister group to Mollusca. Within Mollusca the Bayesian tree supports basal position of Aculifera

(molluscs with no conch or shell; *Cheatopleura apiculata*). Within Conchifera (shell-bearing molluscs), Cephalopoda (*Sepia officinalis*) are basally positioned to Bivalvia and Gastropoda. This result is in agreement with two well-sampled phylogenetic analysis of transcriptome data published in Nature and highlights the good performance of OMA standalone pipeline in phylogenetic analysis (Smith et al. 2011 and Kocot et al. 2011).

The OMA Bayesian tree, confirms previous findings (Egger et al. 2015; Laumer et al. 2015a), and supports the monophyly of Platyhelminthes (flatworms), with Catenulida (*Catenula lemnae*) the most basally position flatworms, sister group to Rhabditophora (*Macrostomum lignano+ Microdalyellia schmiditi+ Echinoplana celerrima+ Monocelis sp. + Schmidtea mediterranea*). The tree support Macrostomorpha as sister group of all other rhabditophoran orders, as previously shown (Egger et al. 2015; Laumer et al. 2015a). Within other rhabditophorans (Trepaxonemata) the tree supports Policlatida (*Echinoplana celerrima;* in the absence of any Lecitoepithiliata species) sister group to Euneophora (*Microdalyellia schmiditi + Monocelis sp. + Schmidtea mediterranea*). The more controversial is the placement of Proseriata (*Monocelis sp.*) closer to Acentrosomata (Triclada, Bothrioplanida and Neodermata; here represented as just a single taxa *Schmidtea mediterranea*) than to Rhabdocoela (*Microdalyellia schmiditi*), which was first suggested by as part of the order Seriata (Proseriata, Triclada, Bothrioplanida and Neodermata) by authors (Laumer at al. 2014; Martın-Duran and Egger 2012) and considered that Rhabdocoela as the earliest-diverging branch of Euneoophora (Proseriata, Rhabdocoela, Triclada, Bothrioplanida and Neodermata). This is in agreement with recent large platyhelminth molecular phylogenies featuring 47 flatworm species (Egger et al. 2015 and Laumer et al. 2015a).

The species trees constructed using the maximum likelihood (ML) method implemented in RaXML yields similar topology, and supports the monophyly of platyhelminthes, molluscs and annelids (see Figure 4.8). However, the maximum likelihood trees places rotifers (*Adineta ricciae*, *Brachionus plicatilis*) as a sister group of nematodes. We obtained similar results in Egger et al. 2015, where ML and fast evolving genes analysis grouped Rotifera with Nematoda. This is in disagreement with the Bayesian tree calculation, which used the CAT model, and is most likely

a result of a Long Branch Attraction artefact in the RaXML tree. Importantly, CAT models have been shown to successfully counteract Long Branch Attraction artefact (Lartillot et al. 2013). Both Baysian and ML phylogenies is in agreement with previous findings. The phylogenies support previously published phylogenies of Mollusca (Struck et al. 2011; Kocot et al. 2011) and Platyhelminthes, and confirm the good performance of OMA standalone pipeline.

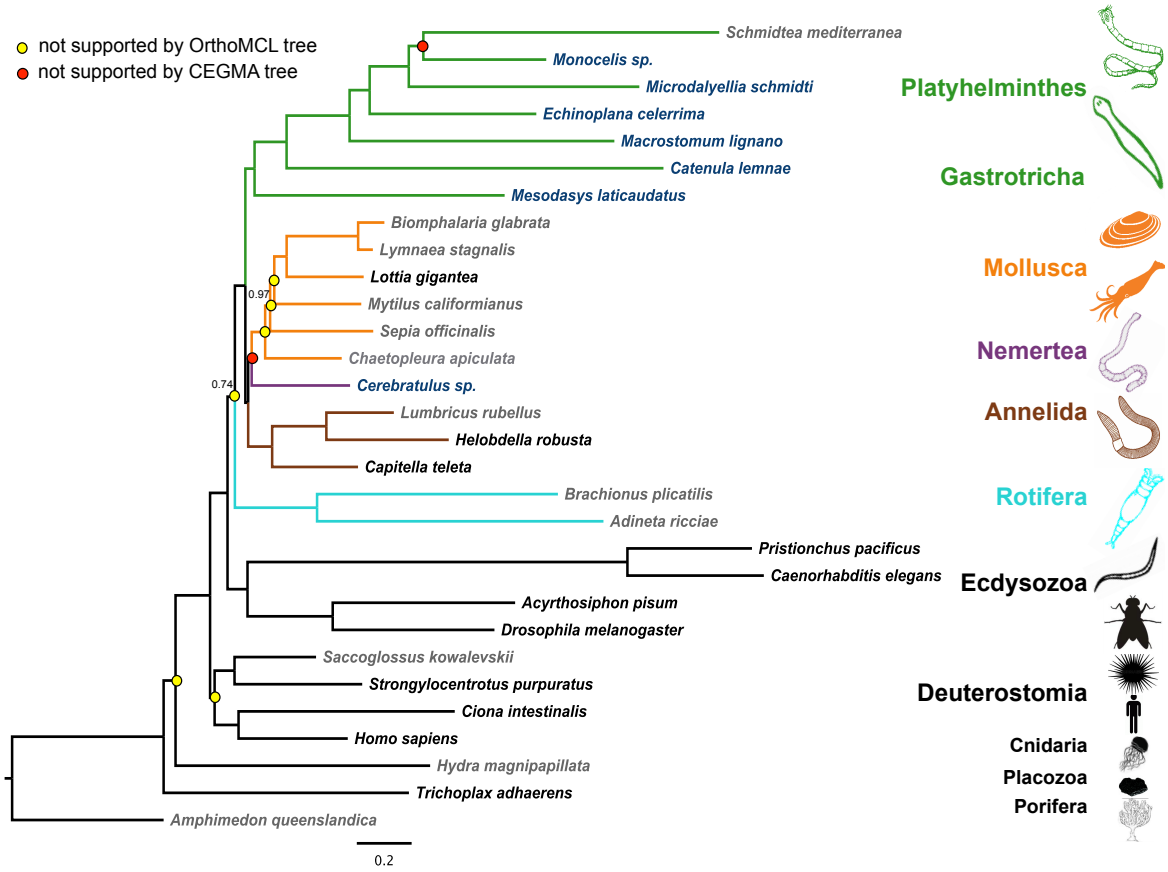


Figure 4.7. The Bayesian phylogeny calculated with OMA pipeline using CAT GTR Γ model in PhyloBayes. Posterior probabilities (PP) lower then 1 are indicated on the nodes. Nodes that are not supported by the Bayesian phylogenies calculated with CEGMA and OrthoMCL pipelines are highlighted using collared dots.

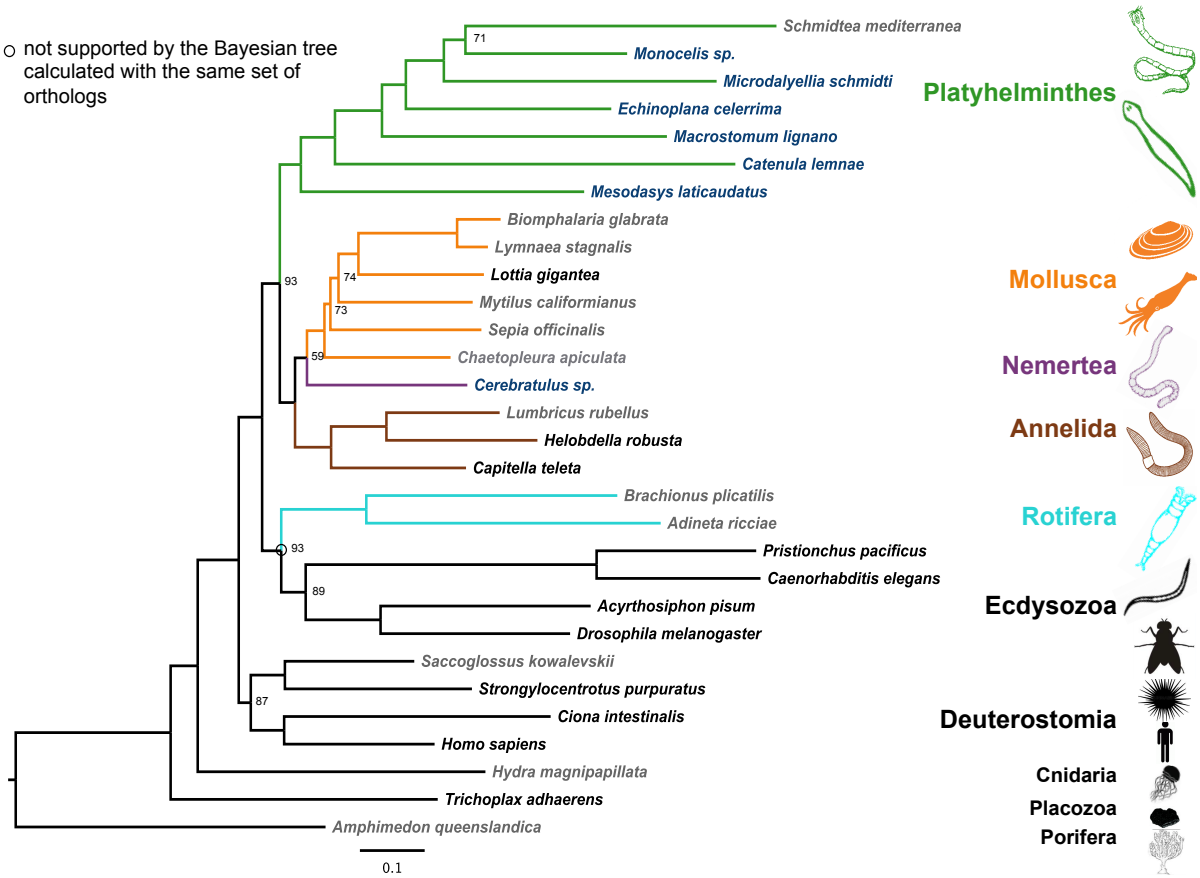


Figure 4.8. The ML phylogeny calculated with OMA pipeline in RaXML. Bootstrap support lower then 100 is indicated on the nodes. Nodes that are not supported by the Bayesian phylogenies calculated with OMA pipeline with the same dataset are highlighted using transparent dots.

4.3.4.1 The analysis of CEGMA dataset

The Bayesian tree calculated with the CEGMA pipeline, yields a same topology as OMA tree apart from the placement of two taxa, and contains more weekly supported nodes (with pp < 1) (see Figure 4.9). The tree represents a different placement of *Monocelis* sp. (Proseriata), *Cerebratulus* sp. (Nemertea). The CEGMA tree supports a basal position of Proseriata (*Monocelis* sp) relative to Rhabdocoela (*Microdalyellia schmiditi*) and Acentrosomata (*Schmidtea mediterranea*), and contradicts the results obtained by (Laumer et al. 2015a; Laumer et al. 2015b; Egger et al. 2015). CEGMA Bayesian tree places Nemertea as a sister group to Platyhelminthes instead of to Mollusca (see Figure 4.9, different nodes marked with blue and black dots). It is difficult to conclude about the performance of CEGMA pipeline based on the position of Nemertea, as both positions have been suggested in the literature and the correct placement of *Cerebratulus sp.* is still debated.

The tree was calculated based on fewer amino acid positions, but with the same minimum alignment density threshold, therefore contains less informative sites, and may be considered less likely to recover a correct tree. Previous analysis by Egger et al. 2015 show that the phylogeny constructed based on fast evolving genes supports the placement of Rhabdocoela sister to Acentrosomata. This observation lets us suspect that CEGMA dataset may contain more fast evolving genes than OMA dataset. Furthermore, missing data in the CEGMA alignment may explain a different placement of *Monocelis sp.* and *Cerabratulus sp., Monocelis sp.* has 285,884 amino acids that are included in the superalignment constructed with OMA for both species, whereas only 64,793 amino acids are included in the superalignment constructed with CEGMA. *Cerabratulus sp.* has 279,897 amino acids in the OMA superalignment and only 62,160 amino acids in the CEGMA alignment. Thus, over 200,000 more amino acid positions are present in the OMA alignment for these two species.

The tree calculated based on the CEGMA orthology set using the Maximum Likelihood (ML) method is not consistent with the literature (Laumer et al. 2015a; Laumer et al. 2015b; Egger et al. 2015; Smith et al. 2011 and Kocot et al. 2011) and the Bayesian tree, and yields different placement of Nemertea, Rotifera, Cephalopoda, Bivalvia and Gastropoda within Mollusca (see Figure 4.10, marked with transparent dots). Maximum Likelihood method does not use site heterogeneous model, but only uses general time reversible model with 4 gamma categories for substitution rate, thus uses less fitted model. Less complete data as provided gives incorrect tree topology with less fitted model, but provides better-reconstructed tree with better-fitted CAT+GTR+ Γ model.

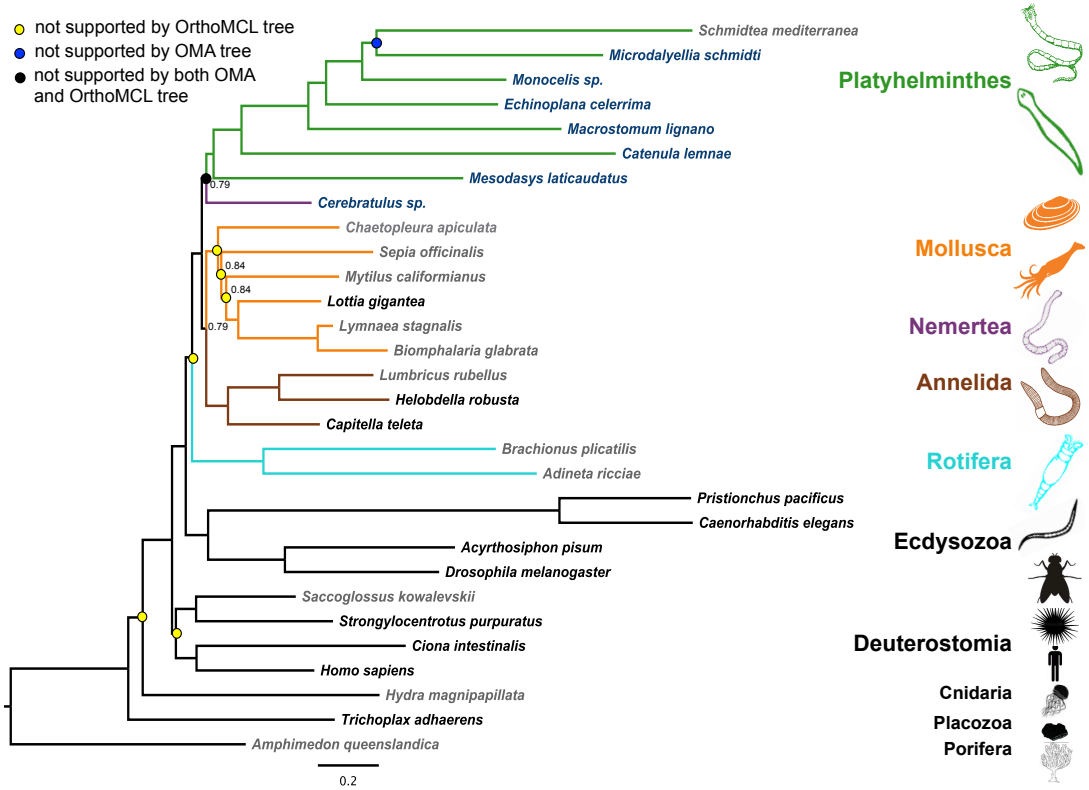


Figure 4.9. The Bayesian phylogeny calculated with CEGMA pipeline using CAT+GTR+Γ model with PhyloBayes. Posterior probabilities (PP) lower then 1 are indicated on the nodes. Nodes that are not supported by the Bayesian phylogenies calculated with OMA and OrthoMCL pipelines are highlighted using collared dots.

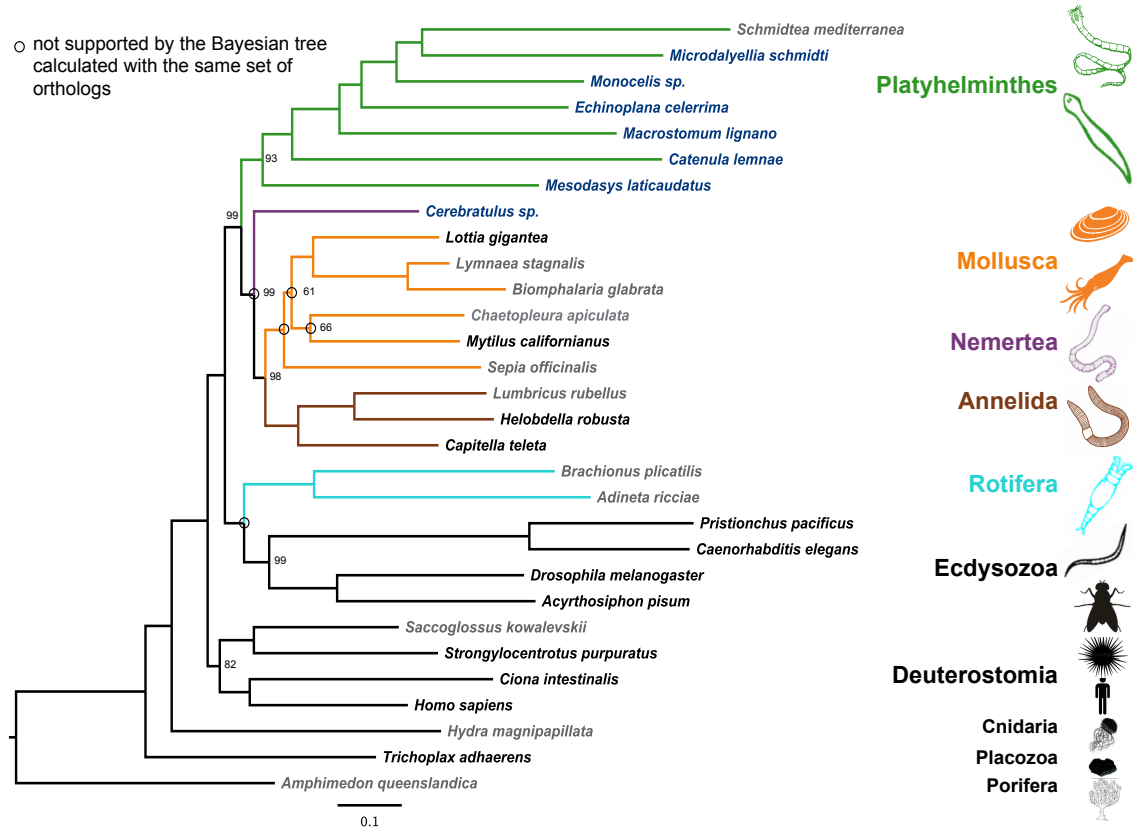


Figure 4.10. The ML phylogeny calculated with CEGMA pipeline in RaXML. Bootstrap support lower then 100 is indicated on the nodes. Nodes that are not supported by the Bayesian phylogenies calculated with CEGMA pipeline with the same dataset are highlighted using transparent dots.

4.3.4.2 The analysis of OrthoMCL dataset

The Bayesian tree calculated with the OrthoMCL orthology pipeline is in disagreement with the current literature, and differs the most from both the OMA and the CEGMA Bayesian trees (see Figure 4.11, differences marked with red, blue and transparent dots). On this tree, Deuterostomes (subtaxon of the Bilateria, where the first opening (the blastopore) becomes the anus; on a tree represented by: *Saccoglossus kowalevskii, Homo sapiens, Strongylocentrotus purpuratus, Ciona intestinalis*) are not a monophyletic group (we will consider the possibility of paraphyletic Deuterostomia in chapter 6 with the dataset which includes more then 4 deuterostomes). Moreover, the organization of Cephalopoda, Bivalvia and Gastropoda within Mollusca is not consistent with two most recent large-scale analysis (Smith et al. 2011 and Kocot et al. 2011), and differs from both the OMA and the CEGMA Bayesian tree. The placement of Rotifera at the base of Platyhelminthes does not support previous large-scale analysis of Lophotrochozoa (Laumer et al. 2015a; Laumer et al. 2015b; Egger et al. 2015; Struck et al. 2014) and Bilateria (Philippe et al. 2011), but is in agreement of previous analysis by Hejnol et al. (2009), suggesting artificial placement of Rotifera using OrthoMCL algorithm.

The Maximum Likelihood (ML) tree calculated with the OrthoMCL pipeline shows most deviations from already published phylogenies of Lophotrochozoa (see Figure 4.12, differences between ML and Bayesian tree are marked with transparent dots) and has a very week bootstrap support (as indicated on the nodes, see Figure 4.12). The ML OrthoMCL tree shows that Platyhelminthes and Ecdysozoa are paraphyletic, which does not support any of the previous findings (Laumer et al. 2015a; Laumer et al. 2015b; Egger et al. 2015; Struck et al. 2014). The ML OrthoMCL tree does not support the order of the taxa within Mollusca (Struck et al. 2011 and Kocot et al. 2011). Both ML and Bayesian OrthoMCL trees are in disagreement with current literature (Laumer et al. 2015a, Laumer et al. 2015b; Egger et al. 2015; Struck et al. 2014; Philippe et al. 2011; Struck et al. 2011; Kocot et al. 2011) and with each other. Both trees were calculated based on the supermatrix with the lowest gene occupancy, supermatrix density and orthology groups recovering the

monophyly of Lophotrochozoa clade less often. This result highlights that supermatrix quality has a major influence on the animal phylogeny reconstruction.

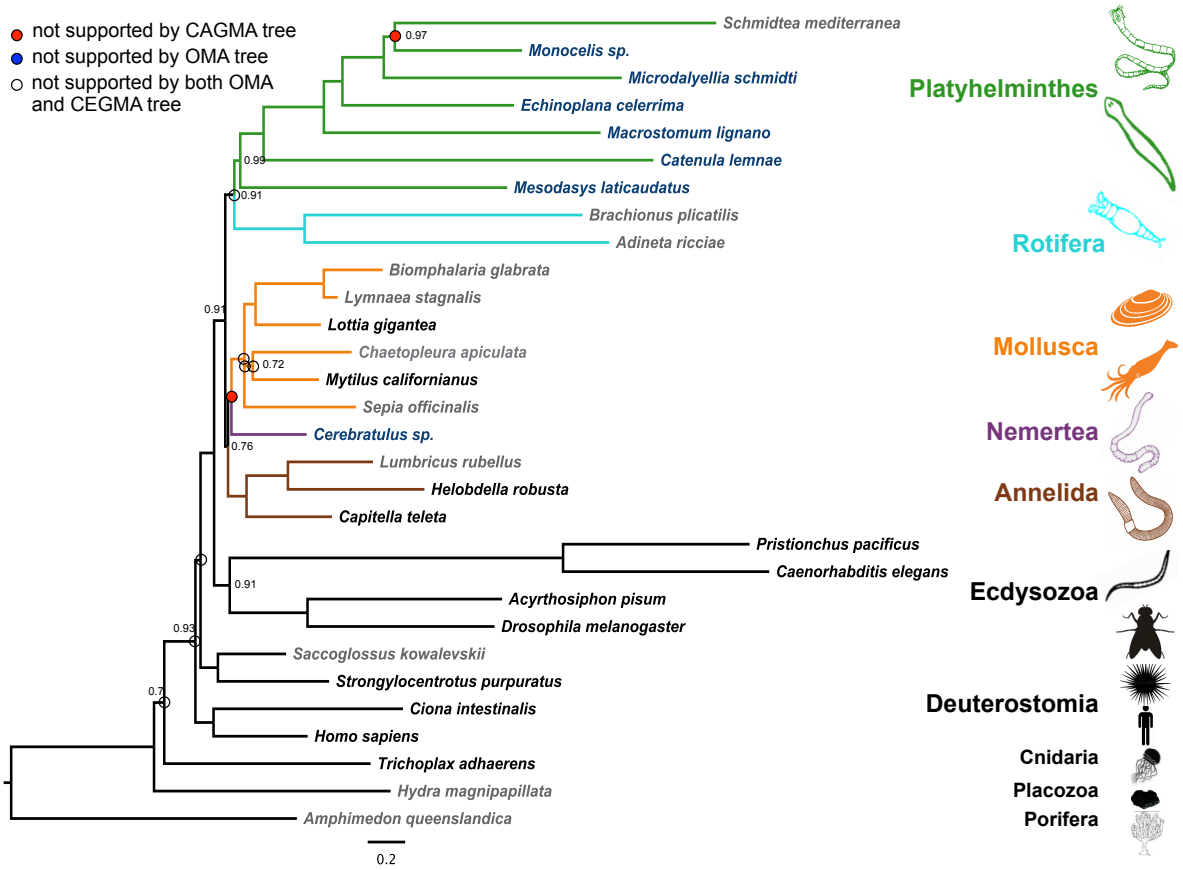


Figure 4.11. The Bayesian phylogeny calculated with OrthoMCL pipeline using CAT+GTR+Γ model with PhyloBayes. Posterior probabilities (PP) lower then 1 are indicated on the nodes. Nodes that are not supported by the Bayesian phylogenies calculated with OMA and CEGMA pipelines are highlighted using collared dots.

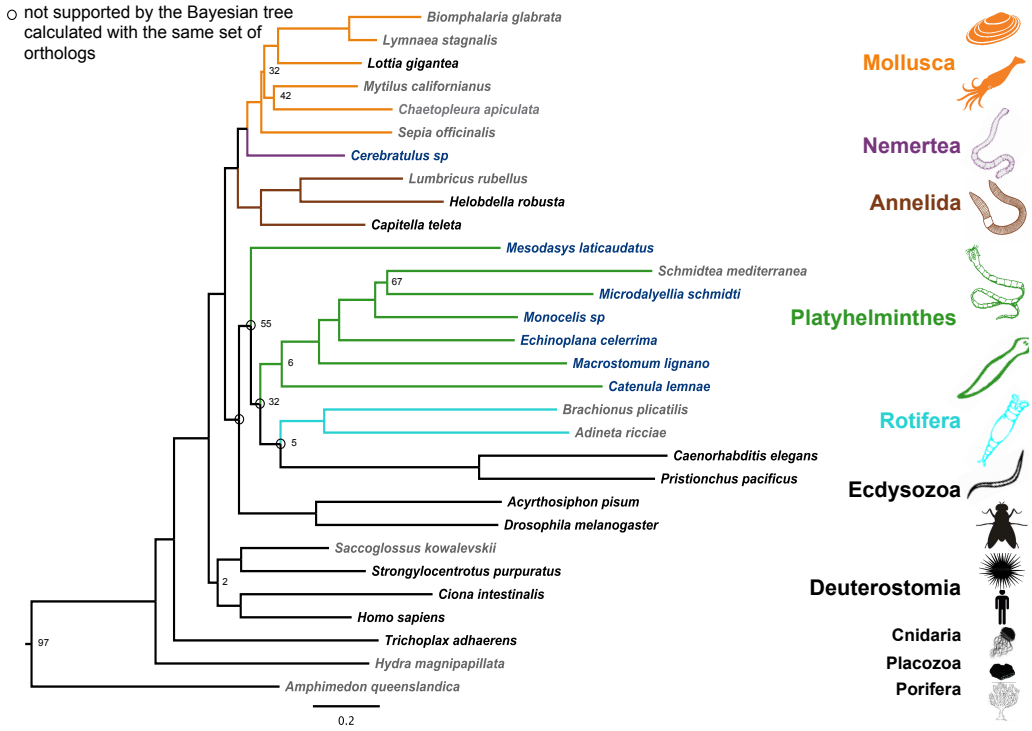


Figure 4.12. The ML phylogeny calculated with OrthoMCL pipeline in RaXML. Bootstrap support lower then 100 is indicated on the nodes. Nodes that are not supported by the Bayesian phylogenies calculated with OrthoMCL pipeline with the same dataset are highlighted using transparent dots.

4.4 Conclusions

The influence of missing data (Roure et al. 2013) and the quality of the data (both for paralogy, orthology predictions, and exogenous contamination) (Philippe et al. 2011; Salichos and Rokas 2011) has a major influence on the reconstruction of phylogenetic trees. Philippe et al. (2011) suggested that automated methods for selection of orthologous genes often introduce ambiguous sequences into superalignments. We gathered genomic and transcriptomic data 30 species (among which 7 new transcriptomes were presented) produced a large and taxonomically complete dataset for resolving difficult phylogenetic question, such as relations within Lophotrochozoa. Using 3 different methods for orthology inference we obtained 3 large phylogenetic alignments (supermatrices) for phylogenetic analysis. We compared the quality of these supermatrices by analyzing its gene occupancy, density and the consistency in reconstructing the monophyly of animal clades. We show some of the cases, where gene trees calculated based on OMA orthology groups recover the monophyly of animal clades, but gene trees calculated based on CEGMA orthology groups do not recover the monophyly

of animal clades because of orthology miss-assignment. We found OMA standalone pipeline perform best in all of these criteria and propose it as the best method for phylogeny reconstruction. Next, we use supermatrix obtained OMA standalone pipeline to reconstruct the known, but hard to reconstruct clades phylogeny, of Lophotrochozoa using both ML and Bayesian methods. We find both trees to be consistent with recent large-scale analysis (Laumer et al. 2015a; Laumer et al. 2015b; Egger et al. 2015; Struck et al. 2014; Philippe et al. 2011; Struck et al. 2011; Kocot et al. 2011).

Additionally, we performed the same type of phylogenetic analysis using supermatrices obtained with CEGMA and OrthoMCL pipelines. We found that only Bayesian analysis reconstructed the monophyly of Lophotrochozoa, where ML analysis consistently grouped Rotifera with Nematoda as a result of Long Branch Attraction artifact. Both OMA and CEGMA Bayesian trees, as well as OMA ML tree, support the placement of Cephalopoda as a sister group to Bivalvia and Gastropoda on a phylogeny (as previously shown Struck et al. 2011; Kocot et al. 2011), where CEGMA ML tree and both OrthoMCL trees do not support that way of evolution. All the trees, apart from OrthoMCL ML tree, support the monophyly of Platyhelminthes and reconstruct the phylogenetic relations within flatworms. Only OMA and OrthoMCL Bayesian trees, as well as OMA ML tree, support that Rhabdocoela as the earliest-diverging branch of Euneoophora, which is in agreement with latest large scale analysis (Laumer et al. 2015a; Laumer et al. 2015b; Egger et al. 2015). Even thought, the position of Nemertea (ribbon worms) is debated in the literature (Laumer et al. 2015a; Laumer et al. 2015b; Egger et al. 2015; Smith et al. 2014; Philippe et al. 2011; Struck et al. 2011; Kocot et al. 2011), and the placement of this taxa is not supportive for any of the analysed pipelines. However, both OMA and OrthoMCL trees support basal position of Nemertea to Mollusca and support previous results obtained by Hejnol et al. (2009) and by a phylogenetic analysis of EST data by Struck et al. (2008). The majority of our conclusions suggest that OMA Bayesian tree is the most consistent with the current state of knowledge about Lophotrochozoa evolution. Moreover, Bayesian tree inference using CAT+GTR+Γ model, performs better where the dataset is lower quality, which confirms previous findings by Philippe. All this suggests that the OMA standalone pipeline, together with Bayesian tree inference using CAT+GTR+Γ model, is a reliable method to reconstruct orthologous groups and can, thus, be

used as a reliable phylogenetic pipeline to construct supermatrices and to perform large-scale phylogenetic analysis deep-lying nodes in animal phylogeny. Led by the evidence presented here, we will use OMA standalone pipeline, together with Bayesian tree inference using CAT+GTR+Γ model, for reconstructing Metazoa phylogeny in Chapter 6.

Chapter 5

The construction of Metazoa gene family database, involving protein sets from 67 species, including 8 Xenacoelomorpha

5.1 Introduction

While there appears to be a consensus forming that the Xenacoelomorpha constitute a monophyletic group, the phylogenetic position of the Xenacoelomorpha within the Metazoa is still being debated in the literature (Hejnol et al. 2009; Philippe et al. 2011; Telford et al. 2013; Srivastava et al. 2014; Cannon et al. 2016; Rouse et al. 2016). One recent molecular phylogeny featuring Xenacoelomorpha places these problematic worms as a sister group of Ambulacraria within the deuterostomes (Philippe et al. 2011). However, more recent results of phylogenetic reconstruction support a phylogenetic position of the Xenacoelomorpha as the most basal bilaterian, a sister clade to deuterostomes and protostomes (Hejnol et al. 2009; Cannon et al. 2016; Rouse et al. 2016). To determine which of these scenarios is more plausible and to understand the evolution of this controversial phylum, more evidence is needed.

Prompted by previous reports of gene absence, in particular absence of Hox and ParaHox paralogs as well as absence of bilaterian miRNAs in Xenacoelomorpha (Hejnol et al. 2009; Cook et al. 2004; Philippe et al. 2011), we are interested in investigating whether this is a prevalent phenomenon affecting the Xenacoelomorpha and, if so, we wish to examine the correlation between the frequency of gene loss and the evolution of simple morphology of the xenacoelomorph worms. The interpretation of this phenomenon depends on their phylogenetic position. If Xenacoelomorpha are sister group to other Bilateria the absence of genes is interpreted as primary absence. If Xenacoelomorpha are sister of Ambulacraria then they must have lost these bilaterian characters. In the previous PhylomeDB Chapter (see Chapter 3) we investigated the presence of ancestral gene families in Xenacoelomorpha, finding that ancestral Bilateria and Metazoa gene families were present in Xenacoelomorpha at similar

levels to those of other extant bilaterians. Furthermore, we showed that some gene families specific to deuterostomes were also present in Xenacoelomorpha. However, we encountered some problems with our approach. First, PhylomeDB gene families were built based on human seed proteins, meaning that only families that contain human genes could be investigated. Second, a limited number of species were represented in the PhylomeDB database. Third, the PhylomeDB approach only allowed us to investigate losses of whole families, without the ability to follow the evolutionary events occurring within gene families.

Here, we aimed to answer if the increased gene loss on the branches leading to Xenacoelomorpha might be causally linked to the apparent morphological simplification of these worms. We ask if Xenacoelomorpha lost many genes, regardless of their phylogenetic position, either from Xenambulacraria Last Common Ancestor (LCA) or Bilateria Last Common Ancestor (LCA). Moreover, we want to investigate how simplifications in Xencoelomorpha body plan and morphology correlate with the gene content of these animals. The construction of Metazoa gene family database is not only important for inferring the phylogenetic position of Xenacoelomorpha, but also for understanding evolution of genes and gene families within the animal kingdom. We aim to reconstruct the orthology and paralogy relations within gene families across Metazoa and infer the gene content of ancestral animal genomes in order to be able to follow evolutionary events across Metazoa, gene losses (death), gene duplications and *de novo* gene gains (birth), to better understand the animal and gene evolution from the gene centric point of view.

5.1.1 Construction of gene family database

To construct the database and overcome the difficulties that we have encountered using the PhylomeDB database (see Chapter 3), we wanted to extend our analysis, and investigate any type of gene families, not only the ones present in human. We wanted to increase the number of species we analyse, and investigate gene losses and duplications within gene families. To achieve that, we have constructed non-redundant set of protein sets (*i.e.*, the protein sequences associated with every protein-coding gene in all genomes) based on a number of new genomic and transcriptomic resources from *Xenoturbella bocki, Meara stichopi, Symasagittifera roscoffensis, Pseudaphanostoma variabilis* (see Chapter 2) and 23 other metazoan and 7 non-metazoan species available on the OMA export page

and refseq repository (Kersey et al. 2005; Flicek et al. 2010; Kersey et al. 2010; Schneider et al. 2007 (http://cbrg-oma-test.ethz.ch/oma/export/); Pruitt et al. 2007 (ftp://ftp.ncbi.nlm.nih.gov/)). However, instead of starting with data from publicly available databases of orthologs, such as COG/KOG (Tatusov et al. 2003), InParanoid (Östlund et al. 2010), OrthoMCL (Li et al. 2003), EnsemblCompara (Kersey et al. 2010), EggNog (Muller et al. 2010), OrthoDB (Kriventseva et al. 2008), PhylomeDB (Huerta-Cepas et al. 2008), and combining the analysis with sequence data from new Xenacoelomorpha genomic and transcriptomic data, we have created our own database of orthologous sequences and gene families that include 34 species of our choice. Furthermore, based on the analysis presented in this chapter, we extend the species content of our database, to 67 species, including 8 Xenacoelomorpha (*Symsagittifera roscoffensis, Meara stichopi, Isodiametra pulchra, Pseudophanostoma variabilis, Paratomella rubra, Praesagittifera naikaiensis* and *Xenoturbella bocki*).

Here, we aimed to provide more evidence in an attempt to resolve the position of the Xenacoelomorpha phylum. To achieve this, we first described the construction of a relatively small dataset of 34 animal protein sets, including 4 new Xenacoelomorpha protein sets. We applied the OMA standalone software on our dataset (Orthologous Matrix (http://omabrowser.org/standalone/) Altenhoff et al. 2014) to calculate a database of gene families within Metazoa (genes or proteins that are presumed to share common ancestry within the taxonomic range of interest, called HOGs (Henikoff et al. 1997; Altenhoff et al. 2013), the term 'gene family' we will use for groups of genes that originated from single gene at any given last common ancestor (root level)). To estimate the robustness of our database, we investigated the effects of including different animal genomes on the result of our analysis and use this information to guide the construction of a bigger and more complete dataset containing protein sets from 67 species. Furthermore, we showed how the leading phylogeny influences the gene family content in our database and highlight the importance of the correct species tree reconstruction for a comprehensive analysis of gene evolution within the animal kingdom. Moreover, we analysed the content of ancestral Metazoa gene families (which are present in an outgroup to Metazoa and as we show are not dependent on the leading phylogeny) for the presence of Ambulacraria specific gene losses in Xenacoelomorpha. Next, we used inferred ancestral gene content on taxonomic levels of the phylogeny to reconstruct evolutionary events, such as gene losses (death), gene duplications and *de novo* gene gains (birth), on the main branches of our leading animal phylogeny (see 5.1.5 for details). To infer the patterns of gene

birth and death, we compared the presence and the absence of gene families at different taxonomic levels, *e.g.* the set of gene families present in the chordate LCA but absent in the vertebrate LCA. Therefore, we could identify losses of genes occurring on the branch between these two nodes of the tree.

5.1.2 Inferring orthology relations between proteins using the OMA standalone package

OMA standalone is publicly available software (Altenhoff et al. 2014), which compares genes on the basis of evolutionary distance, considers distance inference uncertainty and accounts for differential gene losses (Roth et al. 2008). First, it performs pairwise alignments between every pair of proteins between genomes using the Smith-Waterman algorithm ("all-against-all" phase). The alignment score is calculated using the Pam 224 matrix (Gonnet et al. 1992). For significant alignments, with a score above 85, the alignment score is refined afterwards by searching among all PAM scoring matrixes, to maximize the alignment score. Based on the PAM number of the matrix with the best score, the evolutionary distance is estimated in PAM units (Dessimoz et al. 2006). Next, the mutually closest pairs, within a certain confidence interval, are chosen for further analysis ("stable pairs" phase). Stable pairs are verified by the search in a third party genome for possible paralogy that would indicate differential gene loss. If two homologous sequences can be found in the third genome and if the first of these sequences has the closest evolutionary distance to one member of the stable pair, but the second has the closest evolutionary distance to the other member of the stable pair, these two sequences in the third party genome act as a witness of non-orthology. If no "witnesses of non-orthology" can be found in third party genomes (Dessimoz et al. 2006), the pair becomes verified. Once all pairs are verified, an orthology graph, that represents the orthology relations between all sequences, is constructed (Altenhoff et al. 2012). The edges of this graph correspond to the evolutionary distance between orthologous sequences in PAM units. Based on this graph OMA standalone results in two types of output: the OMA hierarchical groups and the OMA orthology groups. Here, we use the OMA hierarchical groups for the analysis of gene family evolution presented in this chapter, while the application of OMA orthology groups in phylogeny interference is presented in Chapter 3 (in comparison with other methods) and Chapter 6 (the interference of the metazoan phylogeny including 8 species of Xenacoelomorpha).

5.1.3 OMA Hierarchical Orthology Groups - HOGs

In order to obtain gene families from the orthology graph, it has to be fragmented to disconnected unrelated sequences from each other. For the imperfect data, such as real genomic sequences, missing (false negative) and spurious (false positive) orthology predictions exist. Hence, no fragmentation would lead to excessively large clusters of orhology. In OMA standalone this procedure is executed by cutting the orthology graph into subgraphs. Connections with the lowest support (lowest PAM distance) are cut in places where a maximum of two cuts has to be made to divide a graph into subgraphs. This is executed by the randomized minimum cut algorithm (Karger et al. 1995, 1996), by recursively identifying the connected components on the orthology for various taxonomic levels. This procedure results in the grouping of genes (into a gene family) that have descended from a single common ancestral gene (Altenhoff et al. 2013)). Next, based on the leading phylogeny, OMA standalone uses the GETHOGs algorithm to resolve the orthology/paralogy relation within each gene family (Altenhoff et al. 2012). This results in set of taxonomic ranges (each node on a given species tree) and their associated orthologous groups (called hierarchical orthology groups – HOGs). The complete set of genes that have descended from a single common ancestor within a given taxonomic range is, thus, called a hierarchical group (Altenhoff et al. 2012).

5.1.4 Accessing hierarchical group content on different taxonomic levels – familyanalyzer.py

The information about hierarchical groups and their orthology/paralogy relations between family members on different taxonomic levels is written in aorthoxml file format (Schmitt et al. 2008), and can be accessed through the familyanalyzer.py software (developed in Dessimoz group, to which I have contributed; Altenhoff et al. 2015). The familyanalyzer.py program returns an ancestral gene content (with the associated family IDs) that is inferred to be present at the Last Common Ancestor (LCA) of the taxonomic range of interest. For a given node of the leading phylogeny, the ancestral gene is inferred to be present, if the family members can be found in any of the species contributing to the children nodes of the leading phylogeny, or in one of the species contributing to the ancestral node and one of the species contributing to the children node of the phylogeny (*e.g.* for the given taxonomic range Bilateria familyanalizer.py returns the inferred genome content of the bilaterian Last Common Ancestor, if the

gene family members are present in both Deuterostomia and Protostomia, or either Deuterostomia and Protostomia and the non-Bilateria). Additionally, familyanalyzer.py returns a set of associated species and genes for each gene family, which is present at the given taxonomic level (set of orthologs that derive from the same gene).

5.1.5 Comparison of ancestral gene content in the ancestral proteomes allows the reconstruction of evolutionary events

To investigate what happened to the genes on particular branches of the animal phylogeny, we compared the ancestral gene content between the taxonomic levels. We developed customized software that compares the gene family content of two taxonomic levels (compare-levels.pl). Following the branches of the leading phylogeny, we inferred the gene loss (death), if the same gene family was present at the parental taxonomic level but is absent at the child taxonomic level. *De novo* gene gain (birth), if the same gene family was absent at the parental taxonomic level but is present at the child taxonomic level. Gene duplication, if the same gene family present at the parental taxonomic consists of at least two subfamilies (paralogy groups) at the child level.

Hierarchical groups correspond directly to gene trees within a given taxonomical range of interest. To analyse the content of hierarchical groups and to allow us to access the set of genes that have descended from the same common ancestor gene at any given taxonomic level, we used familyanalizer.py software. By comparing gene content at two different taxonomic levels, we identified genes that were lost, duplicated or remained unchanged at any given branch of the phylogeny. The analysis of the gene family database, which we created based on protein sets from 30 species at all taxonomic levels along the animal phylogeny, allowed us to follow evolutionary events affecting gene number within the whole Metazoa. Additionally, we used this approach to reconstruct gene evolutionary events throughout the Metazoa using different evolutionary scenarios for the evolutionary position of Xenacoelomorpha. Without prior knowledge of species tree we considered 3 hypothesis for the evolution within Metazoa (Philippe et al. 2011, Hejnol et al. 2009, Bourlat et al. 2006) and reconstructed the gene evolutionary events assuming this three leading Metazoa phylogenies.

5.1.6 Validating the gene family database and the analysis of gene family content

To validate the HOGs, we investigated how much the randomized minimum cut process influences the content of OMA gene families in our dataset. We show that every run of OMA standalone produces the same number of clade specific gene families, only if inferred with the same leading phylogeny. Furthermore, we compared the impact of different fixed species phylogenies as well as the impact of taxon sampling on content of gene families and the outcome of gene evolutionary events reconstruction. Finally, we employed a customized perl-script, 'hog_parser.pl', to analyse the content of OMA gene families (hierarchical groups at the root level). Our analysis reveals lineage specific gene family losses. The presence of deuterostome specific gene losses in Xenacoelomorpha is suggestive for the phylogenetic position of Xenacoelomorpha (see Methods for family-parser.pl).

5.2 Methods

5.2.1 OMA analysis of gene families

The conducted OMA analysis of gene families took several steps, as indicated in brief below:

i) We combined protein sets from different sources and constructed the dataset of 67 non-redundant proteomes. From that we constructed dataset_1 containing 30 proteomes, dataset_2 by adding 4 Xenacoelomorpha proteomes, and full dataset containing 67 proteomes, which includes 8 Xenacoelomorpha proteomes

ii) We processed each dataset using OMA standalone 0.99w software (http://omabrowser.org) with 3 different leading phylogenies

iii) We quantitatively analysed the presence of gene families in major Metazoa phyla using hog_parser.pl

iv) We quantitatively analysed the presence of gene families on each of the taxonomic levels on the leading phylogeny using family_analizer.py

v) We compared each of the pattern of presence and absence of gene families between all taxonomic levels and infer gene evolutionary events such as duplications, gene losses and *de novo* gene creations using compare_levels.pl

5.2.2 Proteome dataset construction

Eight Xenacoelomorpha genomes, *Symsagittiferaro scoffensis, Meara stichopi, Nemertoderma westbladi, Xenoturbella bocki, Pseudophanostoma variabilis, Paratomella rubra, Praesagittifera*

naikaiensis and *Pseudophanostoma variabilis,* were assembled from shotgun reads, using the SOAPdenovo2 assembler as described in Chapter 2 (Luo et al. 2012). Protein sequences were predicted from the Xenacoelomorpha genome assemblies using the GeneScan (Burge et al. 1998). Additionally, 8 Xenacoelomorpha transcriptomes were assembled using the Trinity *de novo* transcriptome assembly software pipeline. Open Reading Frames (ORF) were predicted using the TransDecoder (http://transdecoder.sourceforge.net/). For all peptide datasets cd-hit was used to reduce redundancy by clustering sequences with a global sequence identity of >95%. All subsequent analyses, including the phylogenetic analyses were based on amino acid sequences.

5.2.3 Hierarchical groups processing

The proteomes were placed in a DB folder of OMA standalone 0.99w. The computations were performed on a CS cluster (http://www.cs.ucl.ac.uk/home/) using 500 cores in parallel with the customized bash script. In brief, the OMA algorithm first computes all against-all sequence alignments using full dynamic programming. From these, potential orthologs (‘‘stable pairs’’) are selected based n evolutionary distances and considering inference uncertainty. In a verification step, the algorithm identifies pseudo-orthologs arising through differential gene loss. The resulting ‘‘verified pairs’’ are used to construct the orthology graph for the hierarchical groups as described in Altenhoff et al 2015. From the orthology graph hierarchical groups were inferred with the GETHOGs algorithm and written to the orthoxml file.

The orthoxml file was then parsed using customized sizeFA.pl script and the distribution of the size was plotted in MatLabR2014b using bar.m script. The families from each run were first analysed using familyanaliser.py script at the LUCA level (Last Universal Common Ancestor) (cooperation with Adrian Altenhoff) with the option to propagate top, and the output was analysed using mappgenenames.pl perl script. The taxonomic distance within the family members was analysed using readFA.pl script. The distance between the species was measuredas themaximum number of nodes between two most distantly related species on a leading phylogeny.The presence of gene family members in established taxonomic clades was analysed using hog_parser.pl script. The families with certain patterns of presence and absence in Xenacoelomorpha, Deuterostomia, Protostomia, Ambulacraria, Chordata, Lophotrochozoa and Ecdysozoa were parsed and counted.

5.2.4 Reconstruction of evolutionary events

Each HierarchicalGroups.orthoxml file containing gene families from singe an OMA standalone run was analysed using familyanalyser.pl on every taxonomic level. For every branch of the leading phylogeny, two corresponding levels were compared using compare_levels.pl customized script. If the family 1 was present at both levels gene was ranked as identical. If the family 1 was present at the child level, but absent at the parental level gene was ranked as new. If the family 1 was present at the parental level, but absent at the child level gene was ranked as lost. If the family 1 was present at the parental level, but genes 1.1a and 1.1b were present at the child levels this two genes were ranked as coming from duplication.

5.2.5 GO annotations

First, we assigned GO annotations to genes that are members of Bilateria (http://www.nature.com/ng/journal/v25/n1/full/ng0500_25.html). We used OMA cliques of orthologs to propagate GO annotations among the members of a clique: when one of the orthologs in the respective OMA clique had a GO annotation based either on experimental evidence (GO evidence codes EXP, IDA, IPI, IMP, IGI, IEP) or evidence based on high-quality computational annotations (http://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002533), we propagated the annotation to the OMA group itself, and thereby to the members of Bilateria (cooperation with NivesSkunca). For each gene and orthology group the family ID, gene ID, GO term and support was written to the TSV file.

5.3 Results

5.3.1 Metazoa gene family database

5.3.1.1 Basic characteristics of our gene family database

We first gathered 30 proteomes (dataset_1), which included 959,594 genes total from various resources (NCBI refseq, OMA standalone, Xenocoelomorpha Genome Project 2014) and constructed the orthology database of metazoan gene families using OMA standalone (Altenhoff et al. 2015). The species included 13 Protostomes, 8 Deuterostomes 3 non-bilateria animals (basal Metazoa), closest

living relatives of the animals (*Monosigabrevicollis*, basal Opisthokonta), 1 fungus, 2 plants, 1 Protista and 1 Bacteria (see Figure 5.1). We estimated the evolutionary distance between every pair of genes across proteomes (PAM distance, 882,836,884,836 pairs of genes). Next, based on literature we constructed the leading phylogeny (see Figure 5.1), we used OMA standalone minimum cut algorithm to group genes into families and inferred the paralogy/orthology relation within them (OMA Hierarchical groups).

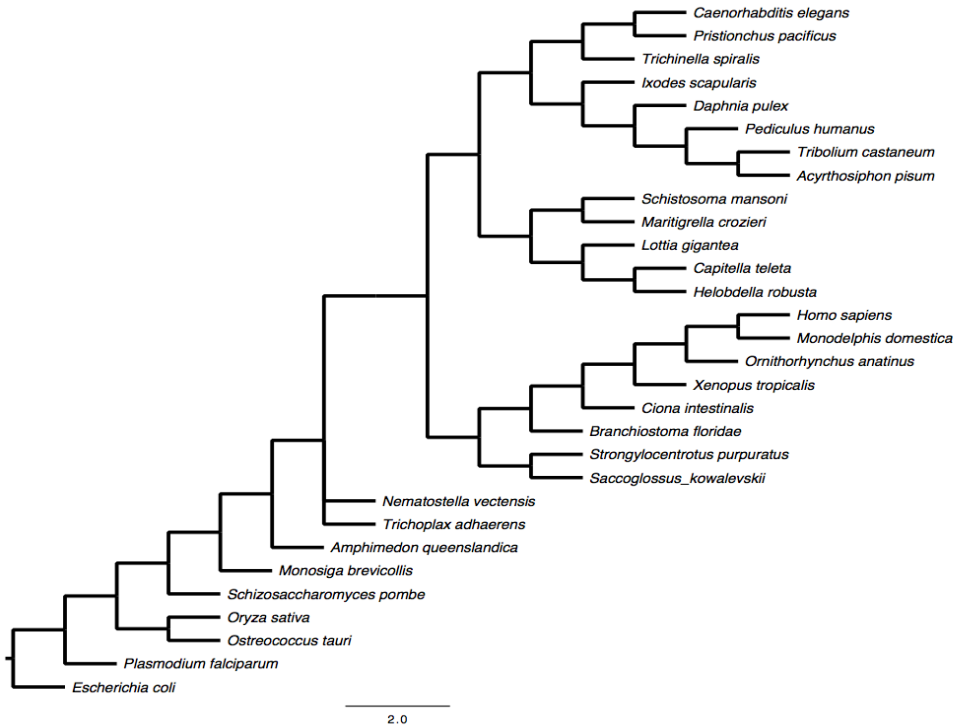


Figure 5.1 The dendrogram representing the phylogenetic relation between 30 species in Metazoa gene family database (dataset_1) used as a leading phylogeny to infer paralogy/ orthology relation within gene families. The species included 13 Protostomes, 8 Deuterostomes 3 non-bilateria animals (basal metazoa), closest living relatives of the animals (*Monosigabrevicollis*, basal Opisthokonta), 1 fungus, 2 plants, 1 Protista and 1 Bacteria.

We obtained 30,932 families with the average size of 8.2 genes. The majority of the families were had fewer then 5 members (median = 4.1), however the biggest families contained up to 500 gene members (see Figure 5.2). Concerned by the randomness of the family grouping process, which is a result of step in OMA standalone algorithm where the orthology graph is cut, we repeated the family grouping 10 times. We found that the number of the families varies from 30,911 to 30,953, with the mean value of 30,923.4 and a standard deviation of 35.2. We compared the content of these gene families using CompareContent.pl perl script and found that on average 0.9% percept of these families have different content of at least one member.

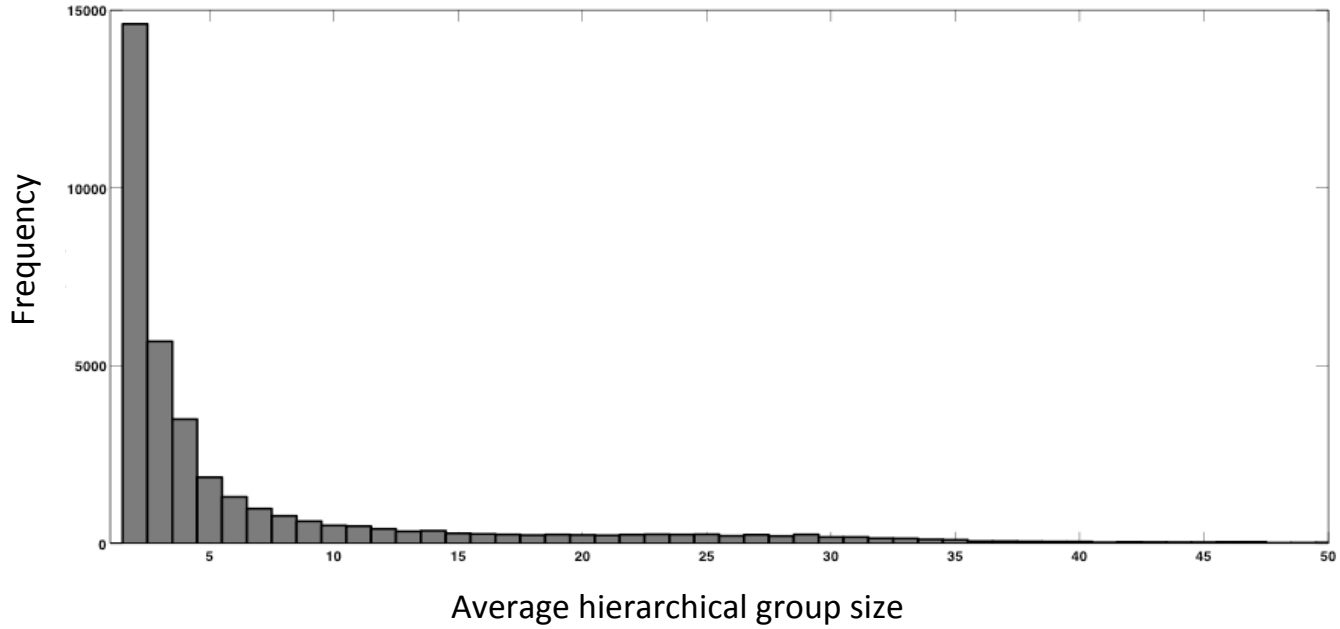


Figure 5.2 The distribution of average family size for 30,932 families calculated for dataset 1. Gene families were calculated with OMA standalone 0.99x with the leading phylogeny from Figure 5.1.

5.3.2 Ancestral Metazoa gene families in our OMA standalone database

Next, we analysed the subset of 13,878 families in our database, which were already present in Metazoa Last Common Ancestor (named Metazoa ancestral families). We found, that these families have a bigger average size of 18.2, and conclude that they are on average bigger than clade specific gene families which appear later on a taxonomic level (see Figure 5.3). The distribution of these families has a characteristic peak near 30, which is the number of species we analysed. Our database contains 4,051 families with more than 15 members, where 31.5% of those are present in more than half of the species in the database (out of 30 present in our database). Only, 3,213 families with more than 30 members are present in the database and 38.6% of those are present in more than half of the species.

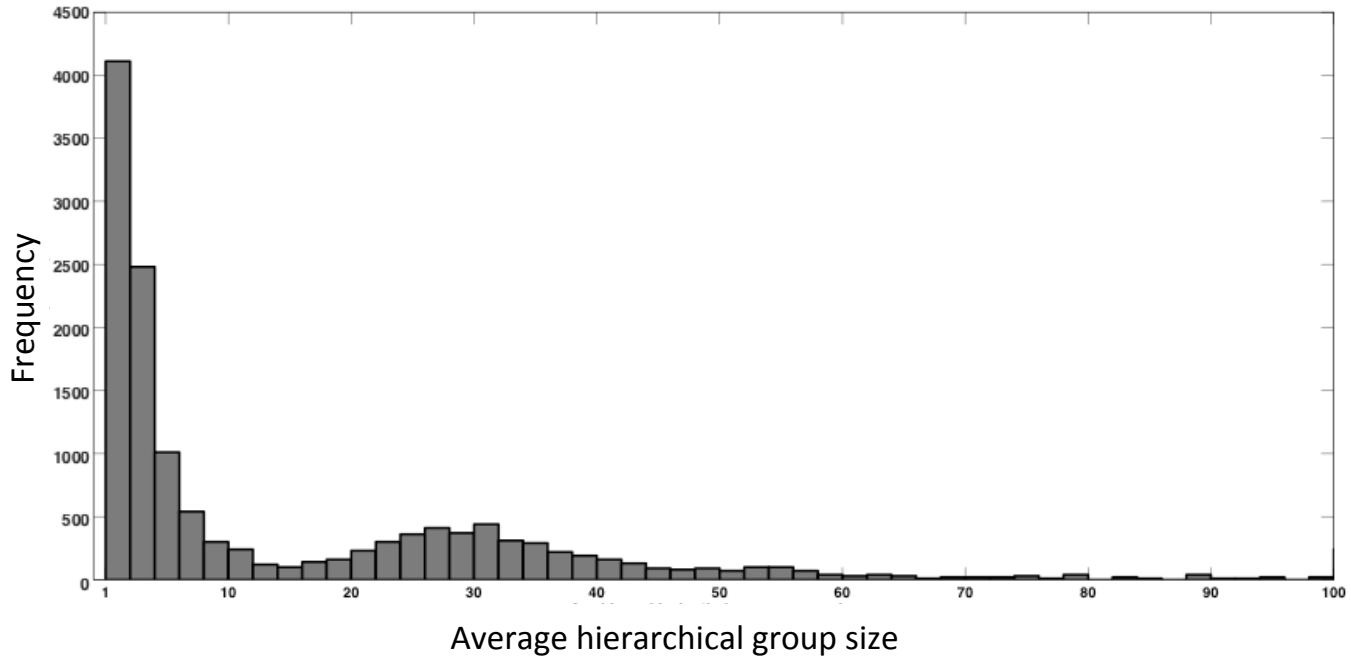


Figure 5.3. The distribution of average family size for 13,878 ancestral to Metazoa (present in Metazoa Last Common Ancestor) families calculated for dataset_1. Gene families were calculated with OMA standalone 0.99x with the leading phylogeny from Figure 5.1.

5.3.3 Metazoa gene family database including 4 new Xenacoelomorpha proteomes

Next we included 4 additional proteomes from acoels *Symsagittifera roscoffensis, Pseudophanostoma variabilis,* the nemertodermatids *Meara stichopi* and xenoturbellid *Xenoturbella bocki* from Xenacoelomorpha clade. Xenacoelomorpha were first placed as a sister group to all other Bilateria on a leading phylogeny according to Hejnol et al. 2009. We found 13,250 gene families present in at least one Xenacoelomorpha, 38.3% (5,076 families) had 5 or less members, had between 6 to 15 members 36.1% of which (4,782 families) 25.6% (3,392) had more than 15 members. We calculated the maximum evolutionary distance between members of Xenacoelomorpha families, by counting maximum number of nodes that between two most distantly related species in a family (see Figure 5.4). We found that the distribution of this distance has two maxima (For example if the family of three genes had members from human, opossum and frog, maximum distance between them is 3 nodes). 32.2% (4,267) of the families are only 3 nodes between each other (which is a distance between human and frog or *Xenoturbella bocki* and *Symsagittifera roscoffensis* on our leading phylogeny) and we could name them clade specific gene families. The other outstanding group of 5,793 (43.7%) families has at maximum 5 to 7 nodes between them (which is a distance between human and *Nematostella vectensis* or *Xenoturbella bocki* and *Caenorhabditis elegans* on our leading phylogeny, Figure 5.1)

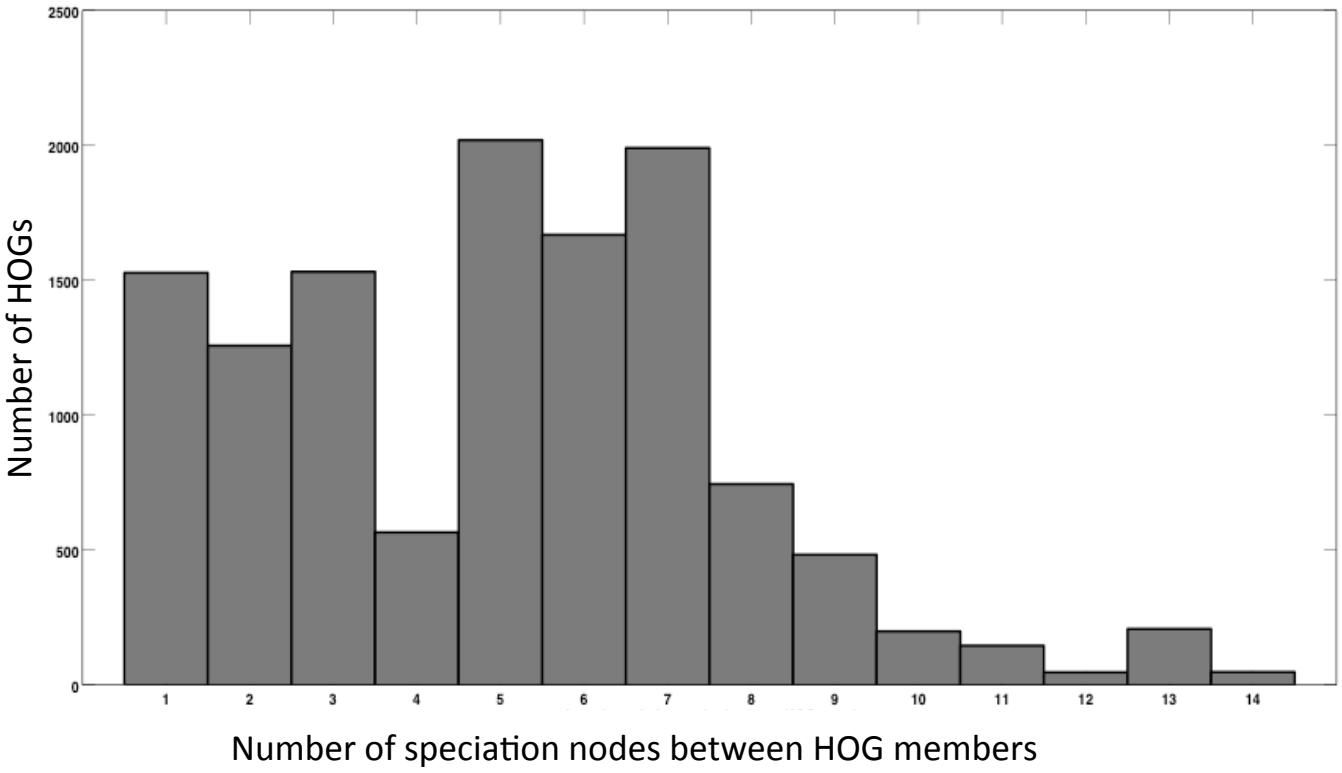


Figure 5.4 Histogram representing the distribution of evolutionary distance between members of 13,250 Xenacoelomorpha families. The distribution of this distance has two maxima. 4,267 (32.2%) clade specific families are only 3 nodes between each other. The other group of 5,793 (43.7%) abundant families has at maximum 5 to 7 nodes between them.

5.3.3.1 The influence of taxa selection on the analysis of gene families

To create the database of Metazoa gene families, we chose 34 taxa motivated by maximum divergence between species in the database with the current knowledge of the animal phylogeny, availability of the data and computational time limitations. Here, we aimed to evaluate the current choice of species, which could influence the result of gene family reconstruction using OMA standalone. We have measured the bias introduced by each genome using a Jackknife method (one of the proteomes was removed from the dataset and the families were recalculated). We measured the number of core gene families present in all major clades of Bilateria (Chordata, Ectodysozoa, Lophotrochozoa, Xenacoelomorpha and Ambulacraria) when all species are present in the dataset. Next, in each run we left out one of the genomes from 34 species dataset and recomputed OMA hierarchical groups. We measured the number of core families when one of the genomes is left out (see Figure 5.6). The difference between the numbers of core families inferred when one of the genomes is left out (signed deviation) and when all proteomes were present was calculated. The difference was the biggest when Xenacoelomorpha and Ambulacraria proteomes were not included in the inference

(marked by green frame on Figure 5.6). Xenacoelomorpha and Ambulacraria proteomes are the most informative for our analysis and introduce the biggest bias to core families inference, even though their content of core orthology groups is of a similar quality then other proteomes in the dataset (see Figure 2.11 in Chapter 2). Xenacoelomorpha and Ambulacraria are not well represented in great numbers in the dataset (4 Xenacoelomorpha species and 2 Ambulacraria species) compared to Chordata, Ecdysozoa and Lophotrochozoa (6, 5 and 8 species). Similarly, we measured the influence of each proteome on the number of families present in at least one 13,250 gene families present in at least one Xenacoelomorpha (see Figure 5.5). The result shows that the biggest influence on the number of Xenacoelomorpha gene families had *Symsagittifera roscoffensis, Pseudophanostoma variabilis, Meara stichopi, Xenoturbella bocki,* therefore xenacoelomorphs itself. The content of the database is not robust to removing any of the Xenacoelomorph from the analysis. The fact that both core and Xenacoelomorpha gene families are highly influenced by the presence of multiple species from Ambulacraria and Xenacoelomorpha led us to extend the current dataset, with the particular attention on adding more proteomes from Ambulacraria and Xenacoelomorpha. We gathered the bigger collection of proteomes (see Chapter 5.3.4), which will help us to better reconstruct the content of Last Common Ancestor of this and other clades.

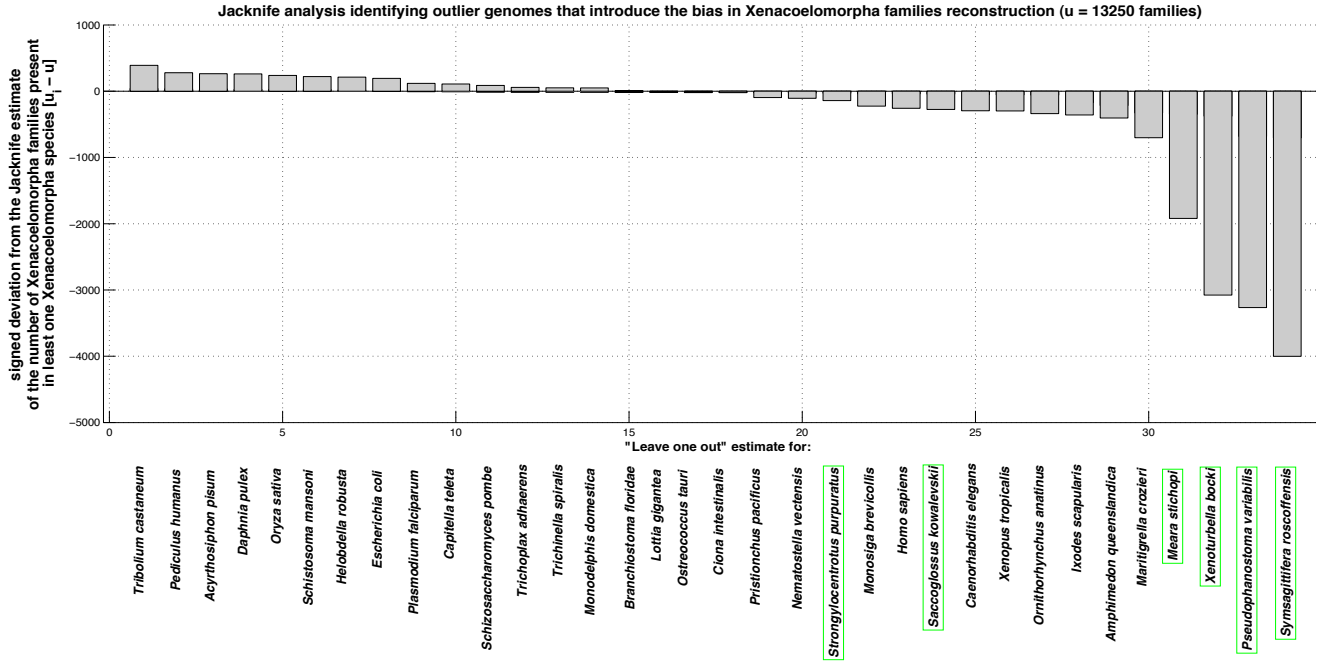


Figure 5.5 The influence of taxa selection on number of core gene families in the database. Xenacoleomorphs *Symsagittifera roscoffensis, Pseudophanostoma variabilis, Meara stichopi, Xenoturbella bocki* and ambulacrarians *Saccoglossus kowalevskii, Strongylocentrotus purpuratus* have the strongest impact on the number of core proteins present in 5 main clades of Bilateria. The removal of *Strongylocentrotus purpuratus or Xenoturbella bocki* reduces the number of families present in Chordata, Ecdysozoa, Lophotrochozoa, Xenacoelomorpha and Ambulacraria by over 20%, meaning these taxa are the most pivotal for the reconstruction of the Last Common Ancestry of both Xenacoelomorpha and Ambulacraria.

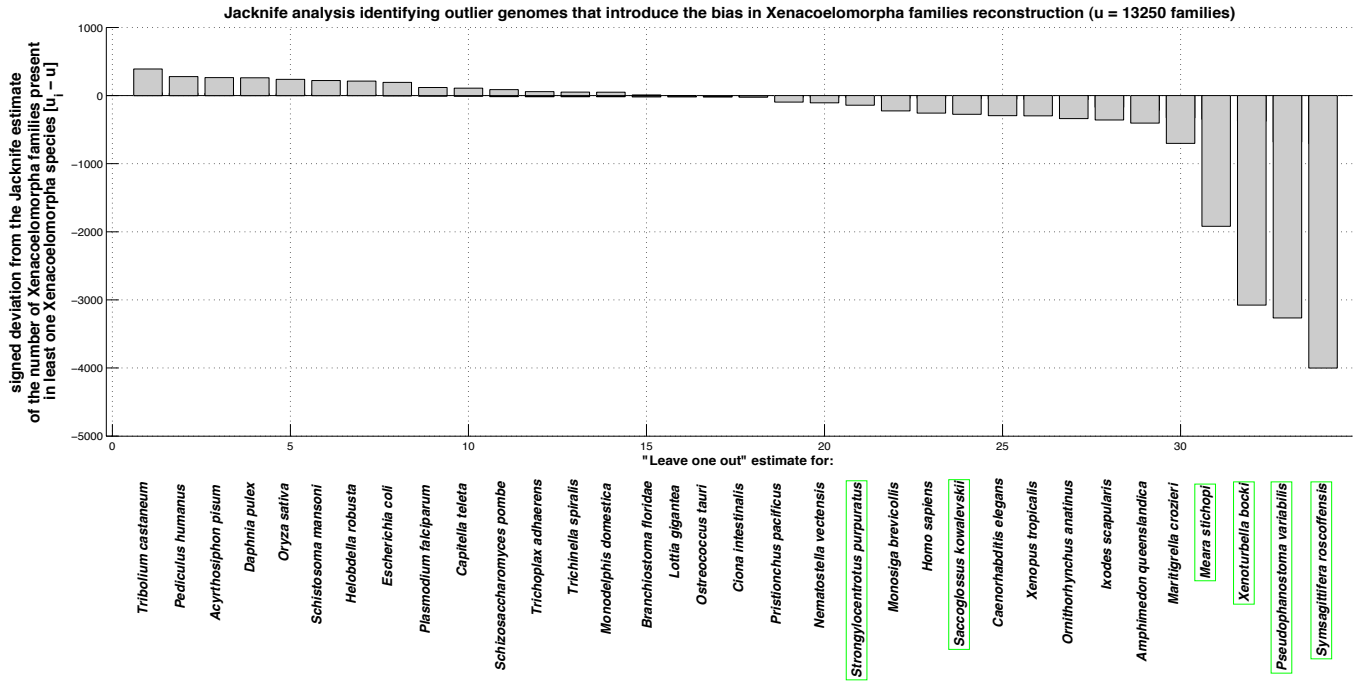


Figure 5.6 The influence of taxa selecton on number of Xenacoleomorpha families in the database. Xenacoleomorphs *Symsagittifera roscoffensis, Pseudophanostoma variabilis, Meara stichopi, Xenoturbella bocki* have the strongest influence on the number of families present in at least one member of this clade. The content of the database is not robust to removing any of the Xenacoelomorph from the analysis.

5.3.3.1 The influence of leading phylogeny on gene family database content

Because the position of Xenacoelomorpha on a tree of life is debated (Telford et al. 2016), we created 3 different scenarios for their evolution within Metazoa (dataset_2): scenario “A” - Xenacoelomorpha are sister group to Ambulacraria (Philippe et al. 2011), scenario “D” - Xenacoelomorpha are basal Deuterostomes (Bourlat et al. 2006), scenario “B” - Xenacoelomorpha are basal Bilateria (Hejnol et al. 2009) (see A, D, B on Figure 5.7). We used these scenarios as a leading phylogeny to infer Metazoa gene families out of 34 proteomes with OMA standalone. The total number of the families changed only by 0.8% (from 37,244 to 36,946) between scenario A and B and by only by 0.6%(from 37,244 to 37,015) between scenario A and D.

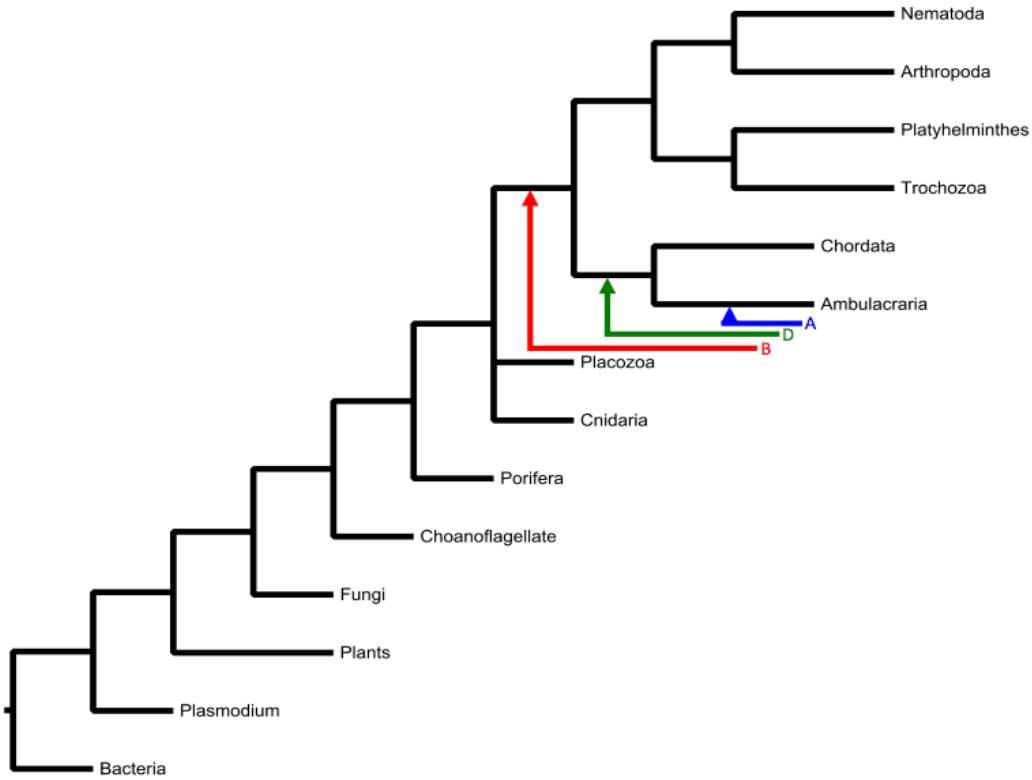


Figure 5.7 Three different positions of Xenacoelomorpha on a tree of life used as a leading phylogeny in the inference of gene family database (dataset_2); (A) Xenacoelomorpha are sister group to Ambulacraria indicated in blue (Philippe et al. 2011). (D) Xenacoelomorpha are basal Deuterostomes indicated in green (Bourlat et al. 2006). (B) Xenacoelomorpha are basal Bilateria indicated in red (Hejnol et al. 2009).

To compare how the content of the inferred gene families with OMA standalone differs depending on the phylogenetic position of Xenacoelomorpha on a leading phylogeny, we investigated how many genes overlap between the same gene families. We generated the random subset of 1,000 families that contained at least one xenacoelomorph species from the families calculated with the scenario A. We mapped one of the xenacoelomorph genes to the family calculated with the scenario D, and counted how many genes overlap between two families (apart from the first one). We found that smaller families (<10)

tend to overlap less frequently between phylogenetic scenarios. Bigger families tend to have the same gene content, when calculated with different leading phylogeny (see Figure 5.8). This is not surprising, as it arises from the way OMA algorithm works. The orthology graph is fragmented starting from basal taxonomic levels of leading phylogeny (Altenhoff et al. 2013). Based on this observation, we will use families >10 in the gene evolutionary events reconstruction.

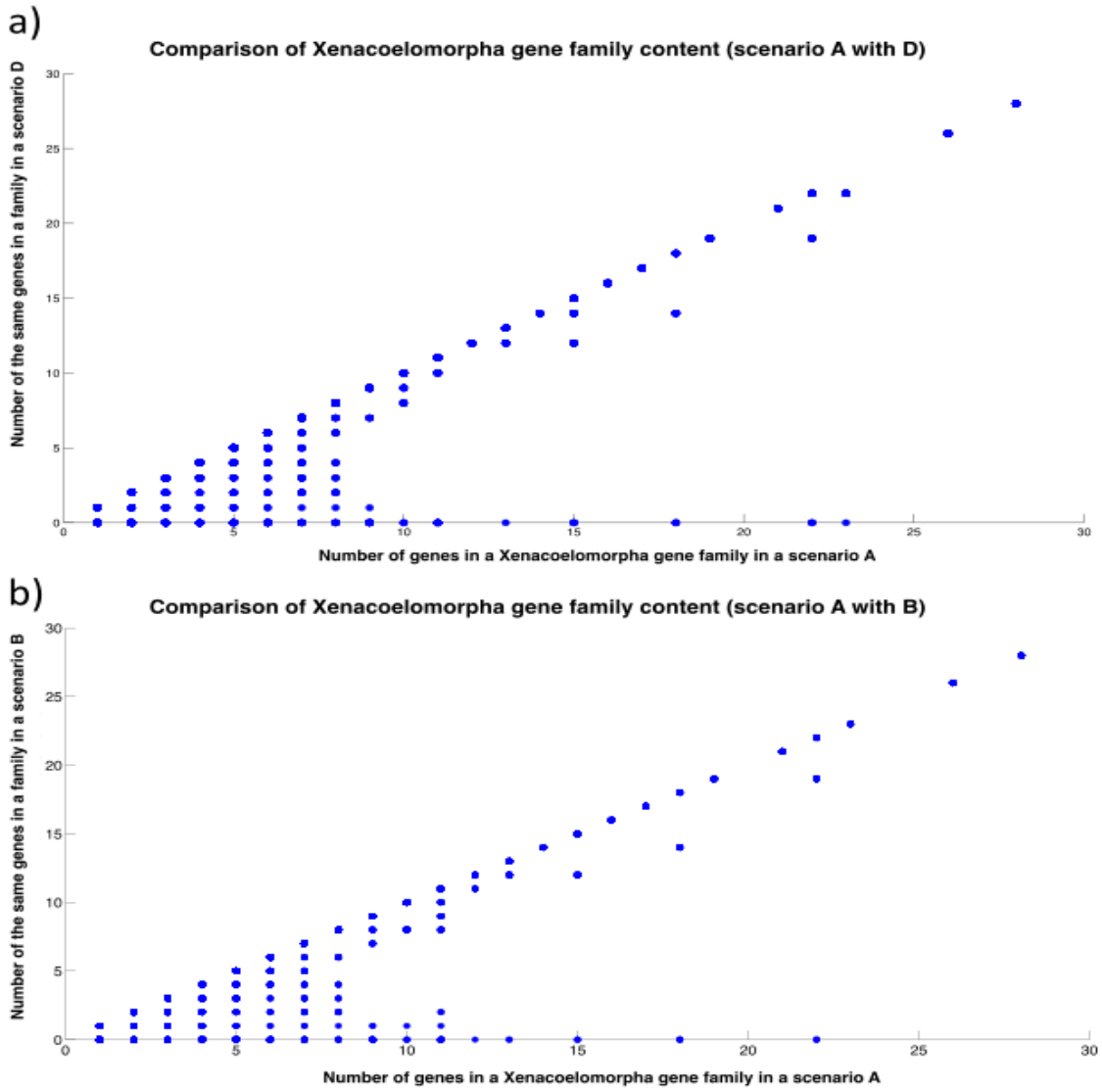


Figure 5.8 Large Xenacoelomorpha families with 10 or more members tend to overlapping gene content more frequently with two different leading phylogenies.a) The dot plot representing the relation between the same genes in a family calculated with a scenario D as with the scenario A, and a family size. b) Relation between the same genes in a family calculated with a scenario B as with the scenario A, and a family size. The dots on a diagonal represent families with the same content. The dots on a) x axis represent the family with no other common members apart from the one gene based on which we mapped both families together. Multiple families are be represented by the same dot if the number of overlapping genes is the same.

Motivated by the observation that bigger families are less dependent on the leading phylogeny, we wanted to investigate, what is the number of families inferred with OMA standalone in each scenario for certain phylogenetic profiles. We designed software (hog_parser.pl) which analyses the content of gene families with aid of OMA standalone, and identifies the families present or absent in 5 well-established monophyletic clades (Xenacoelomorpha, Ambulacraria, Chordata, Lophotrochozoa, Ecdysozoa and non-Bilateria). We investigated how many gene families, with certain pattern of presence and absence of the gene members from each clade, were inferred in a scenario A, B and D (A -Xenacoelomorpha are sister group to Ambulacraria (Philippe et al. 2011), B - Xenacoelomorpha are basal Bilateria (Hejnol et al. 2009), D - Xenacoelomorpha are basal Deuterostomes (Bourlat et al. 2006),).

The number of clade specific gene families is different in each scenario (see Figure 5.9). More Xenambulacraria specific gene families were inferred in a scenario A (Xenacoelomorpha are sister group to Ambulacraria), then in scenario B and D. There were 879 gene families only present in *Xenoturbella* and Ambulacraria in a scenario A (Xenacoelomorpha are sister group to Ambulacraria), but only 408 and 344 gene families respectively with the scenario D and B. Few deuterostome specific genes absent in Xenacoelomorpha (only present in Ambulacraria and Chordata) were inferred in a scenario B (1,974 when Xenacoelomorpha are basal Bilateria) compared to scenario B and D. More deuterostome specific gene families present in Xenacoelomorpha (present in Ambulacraria, Xenacoelomorpha and Chordata) were inferred in a scenario A, than in scenario B and D. The number of clade specific gene families is dependent from the phylogeny, as a consequence of the fact that GETHOGs algorithm in OMA standalone recursively fragments the orthology graph based into families based on the taxonomic levels, which are dependent from leading phylogeny. In the scenario A, Xenambulacraria (Xenacoelomorpha and Ambulacraria) created a taxonomic level, and weak connections to genes from non-Xenambulacraria are cut by GETHOGs, which does not happen in other scenarios. Consequently more Xenambulacraria specific families are inferred with scenario A. Surprisingly, few Deuterostome specific gene families (not present in Xenacoelomorpha) and Ambulacraria specific gene families were inferred with scenario B and D, which could indicate strong similarity to the Xenacoelomorpha genes. This problem requires investigation of individual cases, and is highly influenced by missing sequences as well as imperfection of data, which makes it difficult to interpret. However, almost the same number of gene

families, which were present in the out-group to Bilateria (Xenoturbella, Deuterostomia and Protostomia), is inferred with all three scenarios (A, B and D). Standard deviation for different patterns of presence and absence in established animal clades varies from 0.0 to 2.3. Also the composition of these families was only different in 4.7% of them (as calculated with mappgenenames.pl). We chose these families for further investigation of clade specific gene losses in Section 5.3.7, as they are not dependent from the leading phylogeny and can be informative for the phylogenetic position of Xenacoelomorpha.

						B	D	A	
Xenacoelomorpha	Deuterostomia		Protostomia		outgroup	infrared phylogeny			
	Ambulacraria	Chordata	Lophotrochozoa	Ecdysozoa		Xenacoelomorpha in a basal possition	Xenacoelomorpha sister group to Deuterostomia	Xenacoelomorpha sister group to Ambulacraria	SD
yes	yes	no	no	no	no	344	408	879	292.2
no	yes	no	no	no	no	188	183	256	40.8
yes	yes	yes	no	no	no	1974	1710	1074	462.6
no	yes	yes	no	no	no	71	487	495	242.5
no	yes		yes		no	3128	2785	2787	197.5
yes	yes		yes		no	839	842	842	1.7
yes	yes		yes		yes	784	784	785	0.6
no	yes		yes		yes	1184	1184	1184	0.0
no	no	no	yes	yes	yes	486	484	487	1.5
no	yes	yes	no	no	yes	122	123	122	0.6
yes	no	no	yes	yes	yes	476	472	472	2.3
yes	yes	yes	no	no	yes	47	47	47	0.0
yes	no	yes	no	yes	yes	22	22	22	0.0
yes	no	yes	yes	no	yes	316	317	314	1.5

Figures 5.9 Only ancestral to Metazoa families are independent from the phylogenetic position of Xenacoelomorpha (highlighted in green). The number of clade specific gene families is dependent on the leading phylogeny. More families present in Xenacoelomorpha and Ambulacraria are inferred by OMA in a scenario where the leading phylogeny indicates close relation of this clades.

5.3.4 Gene family evolution over the phylogeny (inferring gene duplication, gains and losses)

We used the families from dataset_1 (30 proteomes) and the given species tree to find single common ancestral genes in the last common ancestor of a given taxonomic range (Altenhoff et al. 2013). The information, which genes within the family descended from a single common ancestor gene, is equivalent with knowing paralogy/orthology relations within the given family under certain leading phylogeny. From that, we were able to infer ancestral gene content on each of the taxonomic levels of the leading phylogeny (for example the common ancestor of *Saccoglossus kowalevskii* and *Strongylocentrotus purpuratus* creates a level on a phylogeny and we can infer a gene content for it). If the orthology group within the family had members in *Saccoglossus kowalevskii* and *Strongylocentrotus purpuratus* we infer this single common ancestral gene to be already present in the Last Common Ancestor of Ambulacraria (in this case Ambulacraria consists only from *Saccoglossus kowalevskii* and *Strongylocentrotus purpuratus*). Additionally, if the orthology group within the family had

members in the out-group to Ambulacraria, and *Saccoglossus kowalevskii* or *Strongylocentrotus purpuratus* we infer this single common ancestral gene to be present in the ancestor of Ambulacraria. We repeat this procedure for every taxonomic level. This allowed us to infer gene content on every taxonomic level of the leading phylogeny, and keep the information about the family origin of the ancestral gene (see Figure 5.1). Next, we compared the gene content between taxonomic levels, on every branch of the tree of lifewithcompare-levels.pl software. We inferred the: gene loss (death) - if the same gene family was present at the parental taxonomic level but is absent at the child taxonomic level; *de novo* gene gain (birth) - if the same gene family was absent at the parental taxonomic level but is present at the child taxonomic level gene duplication - if the same gene family present at the parental taxonomic level consists of at least two subfamilies (paralogy groups) at the child level.

5.3.4.1 Gene content of the ancestor of Bilateria expanded through duplications and *de novo* gene creations and limited number of losses

We inferred 665,111 genes on 57 taxonomic levels on the tree of life, which undergone 375,729 gene events, from which most frequent were losses (220,493 less events, with the average ration 0.33 loss per gene * branch). Less frequent (34% of the events (126,137)) were gene duplication events, and least frequent (8% of all the events (29,099)) were *de novo* gene creations (see Figure 5.10). According to our calculation the Last common Ancestor of Opisthokonta (Metazoa, Fungi, Choanoflagellata) had a small genome content of 4,958 genes which increased almost twice up to 9,887 genes in the Last Common Ancestor of Metazoa (Urmetazoa), through the large number of 5,931 duplications, 538 *de novo* gene creations and only 2,338 losses and with the average evolutionary rate 0.42. We inferred even larger gene content in the ancestor of all Bilateria (19,055), which evolved through 4,764 duplications, 6,776 *de novo* gene creations and only 895 gene losses, with the average evolutionary rate 0.46 (from the Metazoa Last Common Ancestor). Our result indicates that the ancestor of Bilateria already had large repertoire of genes. The ancestor of Bilateria doubled the gene content from Metazoa Last Common Ancestor, through duplications and *de novo* gene creations, while maintaining low gene loss rate.

From the ancestor of Bilateria, deuterostomes lost more genes (5076 losses, 5272 duplications and 2048 *de novo* gene creations), than Protostomes (2882 losses, 5347 duplications and 2097 *de novo* gene creations). Within Protostomia, nematodes (7,928) and platyhelminthes (8,593) lost more genes

compared to Arthropoda (1,835) and Trochozoa (5,229). While within deuterostomes, Ambulacraria lost more genes 7,209 compared to Chordata 3,276, from the deuterostome Last Common Ancestor. However, more duplication events were inferred on a branch leading to Ambulacraria (3,128), compared to Chordata (2,983).

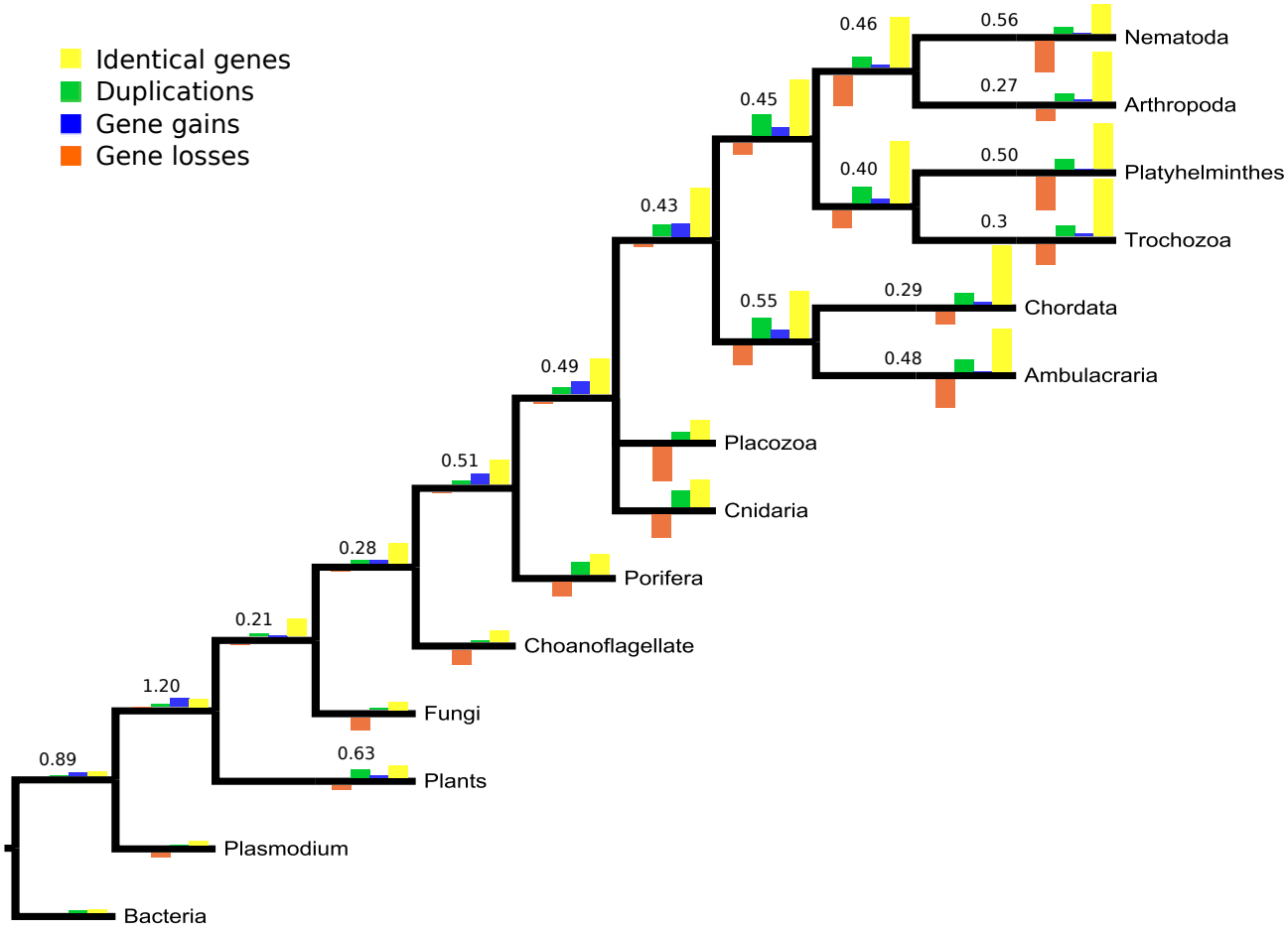


Figure 5.10 Gene family evolution across Metazoa. Following every single gene copy from its origin until now, gene evolutionary events are quantitatively represented on the tree of life. Branch labels represent the rate of evolutionary event per gene (gene duplication, loss, *de novo* gene creation). Bars represent an absolute contribution of duplications (green), losses (red) and *de novo* gene creation (blue). Unchanged genes are notated in yellow.

5.2.4.2 Xenacoelomorpha lost many genes from the ancestor of Bilateria

Next, we placed 4 species of Xenacoelomorpha at the base of Bilateria and inferred the ancestral gene content of animal genomes and gene evolutionary events on the phylogeny according to the scenario B (leading phylogeny with the Xenacoelomorpha at the base of Bilateria) (see Figure 5.11). With this dataset we inferred 436 more genes in the ancestor of Bilateria (19,491), with fewer gene duplications (4,014 compared to 4,764 without Xenacoelomorpha) and gene losses (742 compared to 895 without Xenacoelomorpha), but more *de novo* gene creations (7,219 compared to 6,776 without Xenacoelomorpha). Xenacoelomorpha have 153 genes present before the divergence of Metazoa but not present in Nephrozoa. Have 443 new genes for Bilateria (which are present only in Nephrozoa and Xenacoelomorpha). The ancestor of Nephroza (protostostoms and deuterostomes) was inferred to lose 1,577 genes from the ancestor of Bilateria (Nephrozoa plus Xenacoelomorpha). The ancestor of Nephroza was inferred to have 22,608 genes (3,553 more than in the scenario without Xenacoelomorpha (dataset_1)), in which 2,764 were created *de novo*, 3,089 come from duplication events and 1,577 were lost from the ancestor of Bilateria. The ancestor of Xenacoelomorpha had fewer genes (12,573) then the ancestor of Bilateria and Nephrozoa. We inferred large number of 9,362 losses from the ancestor of Bilateria to Xenacoelomorpha, while gaining only 1,165 and duplicating only 2,135. If Xenacoelomorpha are basal Bilateria, they lost many genes already present in Metazoa (9,362 gene losses), as inferred from the comparison of ancestral gene content of Xenacoelomorpha and Urbilateria ancestral gene content. This suggests that the concept of primarily simple Xenacoelomorpha does not have sense from the genetic point of view, since the gene content of the Xenacoelomorpha is strikingly different from the gene content in Urbilateria.

5.2.4.3 The ancestor of Xenacoelomorpha is genetically more similar to Xenambulacraria ancestor then to Bilateria ancestor.

However, in the scenario A we placed Xenacoelomorpha as a sister group to Ambulacraria and when inferring the ancestral gene content and gene evolutionary events we found several differences (see Figure 5.12). The ancestor of Xenacoelomorpha had more inferred genes (14,250), and more of them were identical with the ancestor of Xenambulacraria (12,100 genes; 84% identical to Xenambulacraria ancestor), compared to ancestor of Bilateria in a scenario B) (9,273 genes 74% identical to Bilateria ancestor). Larger ancestral gene content was apparently a result of previous

duplications, as number of whole Xenacoelomorpha families is similar in both cases (scenario A -13,979; B – 13,945). However, the percentage of identical gene content with the parental level is different. Additionally, fewer genes were inferred to be lost from the ancestor of Xenambulacraria in a scenario A) (8,498) then from the ancestor of Bilateria in a scenario B) (9,362). Fewer gene duplications were inferred from the ancestor of Xenambulacraria in a scenario A) (8,498) then from the ancestor of Bilateria in a scenario B)(1,270 in a scenario A compared to 2,135 in a scenario B), and fewer *de novo* gene creations were inferred from the ancestor of Xenambulacraria in a scenario A) then from the ancestor of Bilateria in a scenario B) (880 in a scenario A compared to 1165 in a scenario B). Moreover, fewer gene duplications, gene losses and *de novo* gene creations were observed when comparing ancestor of Xenacoelomorpha in a scenario D (not shown). This suggests that the ancestral gene content of Xenacoelomorpha is more similar to the ancestor of Xenambulacraria then to the ancestor of Bilateria and Deuterostomia. Additionally, we inferred fewer gene duplications (Figure 5.13)(Xenacoelomorpha - 1270, a- Ambulacraria - 2267, Chordata -3178, Trochozoa - 2741, Platyhelminthes -2590, Nematoda- 1675, Arthropoda – 1845, Arthropoda - 3028), and more gene losses compared to other main Bilateria clades (Xenacoelomorpha- 8498, Ambulacraria - 8043, Chordata - 8041, Trochozoa - 6013, Platyhelminthes - 8804, Nematoda - 8266, Arthropoda - 3028), suggesting molecular reasons for their morphological simplification (Figure 5.13). In hindsight, there were few gene duplication events on a branch leading to Xenacoelomorpha then other major clades, and the content of ancestral Xenacoelomorpha genome is more similar to the ancestor of Xenambulacraria, then to Bilateria. However, the inference of the ancestral gene content is influenced by the quality of the data, and the number of taxa that we used, which is why we decided to extend our analysis by constructing larger dataset of animal proteomes.

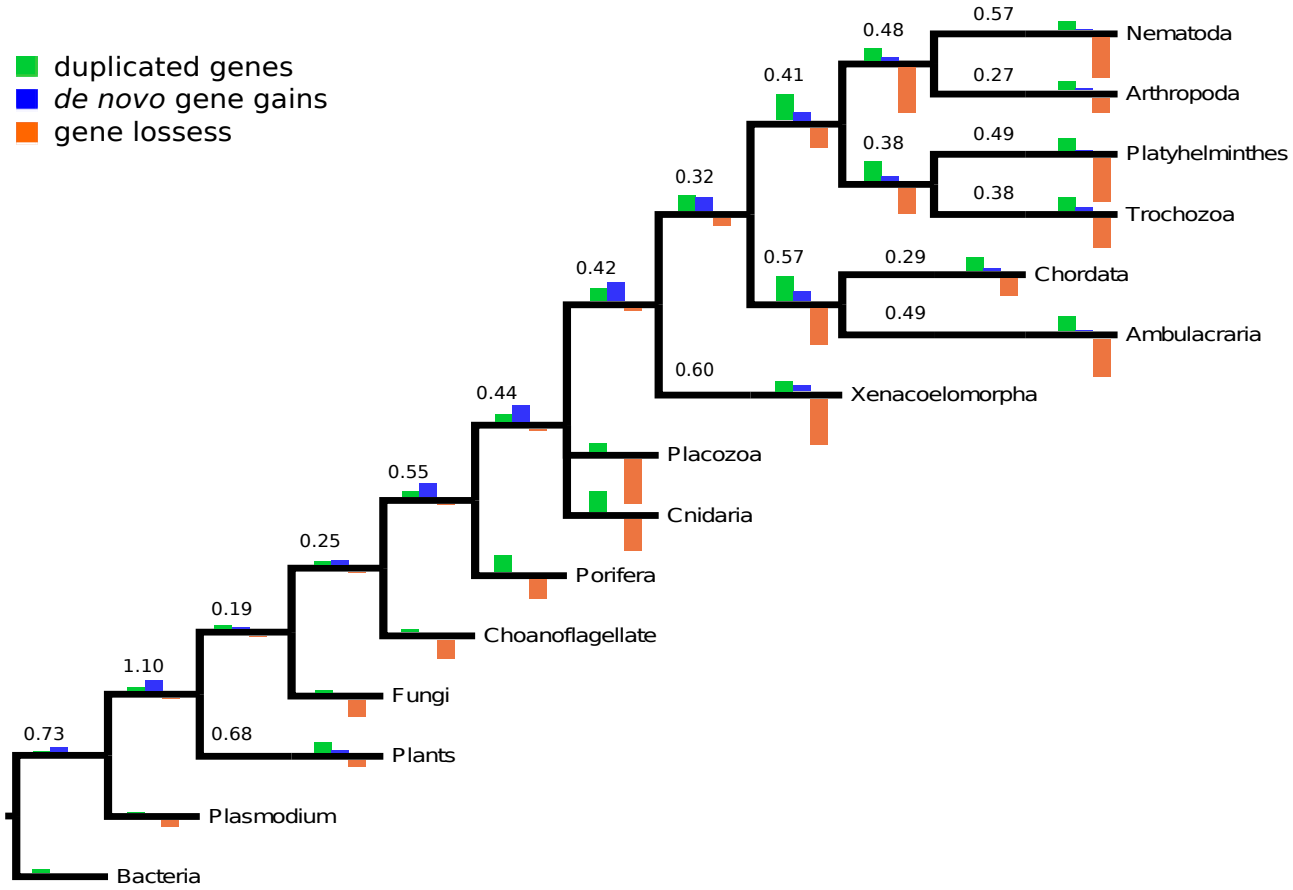


Figure 5.11 Xenacoelomorpha are sister group to Nephrozoa at the base of Bilateria with high evolutionary rate on a branch leading to Xenacoelomorpha Last Common Ancestor (0.6 event per gene). Branch labels represent the rate of evolutionary event per gene (gene duplication, loss, *de novo* gene creation). Bars represent an absolute contribution of duplications (green), losses (red) and *de novo* gene creation (blue).

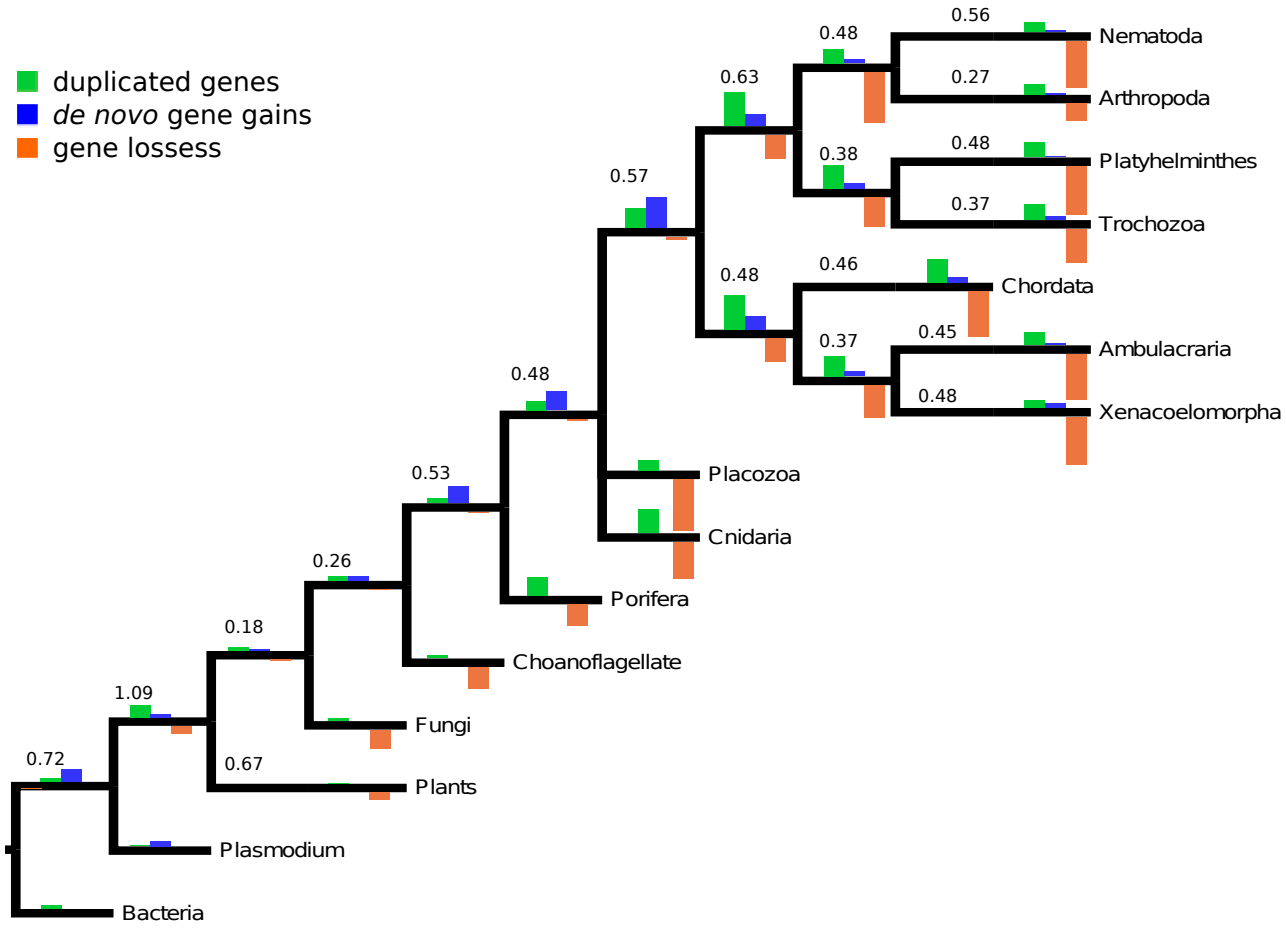


Figure 5.12 Xenacoelomorpha are sister group to Ambulacraria (scenario A) with a similar evolutionary rate as other main clades of Metazoa. Following every single gene copy from its origin until now gene evolutionary events are quantitatively represented on the tree of life. Branch labels represent the rate of evolutionary event per gene (gene duplication, loss, *de novo* gene creation). Bars represent an absolute contribution of duplications (green), losses (red) and *de novo* gene creation (blue).

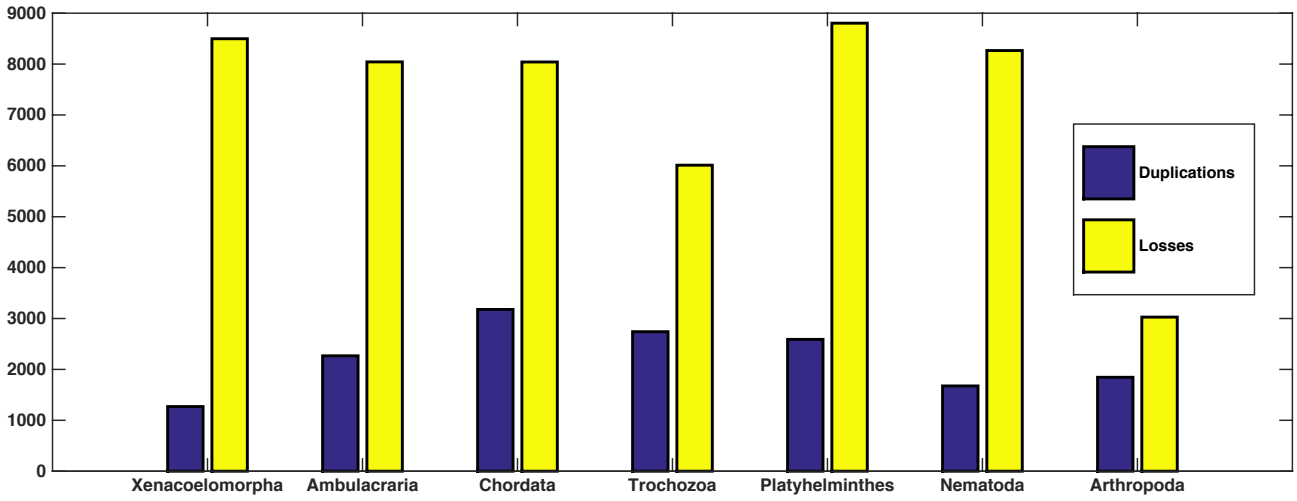


Figure 5.13 Xenacoelomorpha didn't lose more genes then other main animal clades, while maintaining low duplication rate. The number of inferred gene losses and gene duplications on the branches leading to 7 main animal clades. Duplications shown in blue: Xenacoelomorpha - 1270, a- Ambulacraria - 2267, Chordata - 3178, Trochozoa - 2741, Platyhelminthes - 2590, Nematoda - 1675, Arthropoda - 1845: Arthropoda –3028 Losses shown in yellow: Xenacoelomorpha - 8498, Ambulacraria - 8043, Chordata - 8041, Trochozoa - 6013, Platyhelminthes - 8804, Nematoda - 8266, Arthropoda - 3028.

5.3.5 The analysis of extended dataset of 67 species dataset 3

5.3.5.1 Clade specific gene losses

Prompted by the results of the jackknife analysis presented in Section 5.3.3.1 we extended the dataset of proteomes up to 67 species (see Figure 5.17). We inferred a larger set of 2,310,654 gene families, from which 31,690 were present in Xenacoelomorpha. We analysed the subset of 30,176 ancestral gene families (present in the out-group to Metazoa) for the presence of clade specific gene losses using hog_parser.pl perl script (see Figure 5.14). We found that there are more ancestral gene families lost simultaneously in Xenacoelomorpha and Duterostomia then in Protostomia, suggesting closer affinity to this clade. Additionally, we found more gene families simultaneously lost in Xenacoelomorpha and Ambulacraria, than in Xenacoelomorpha and Chordata or Xenacoelomorpha and Lophotrochozoa (see Figure 5.15) (231 with Ambulacraria, 120 with Chordata, 29 with Lophotrochozoa, 294 with Ecdysozoa). However, more ancestral gene families were lost simultaneously with Ecdysozoa (the clade containing nematodes and arthropods) than with other clades.
Our analysis shows that clades, which lost more ancestral gene families, (Ambulacraria and Ecdysozoa) tend to lose more genes simultaneously with the Xenacoelomorpha. To minimize the bias coming from the taxa, which lost genes more frequently, we prepared matrix of gene family presence and absence for the phylogenetic analysis, which is currently being analysed.

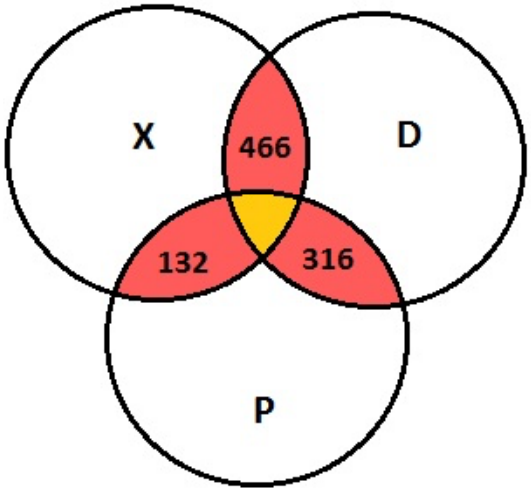


Figure 5.14 Xenacoelomorpha lost more ancestral gene families simultaneously with deuterostomes then protostomes. The number of simultaneous gene losses of the 30,176 ancestral gene families shown in red. X – Xenacoelomorpha, D – Deuterostomia, P – Protostomia.

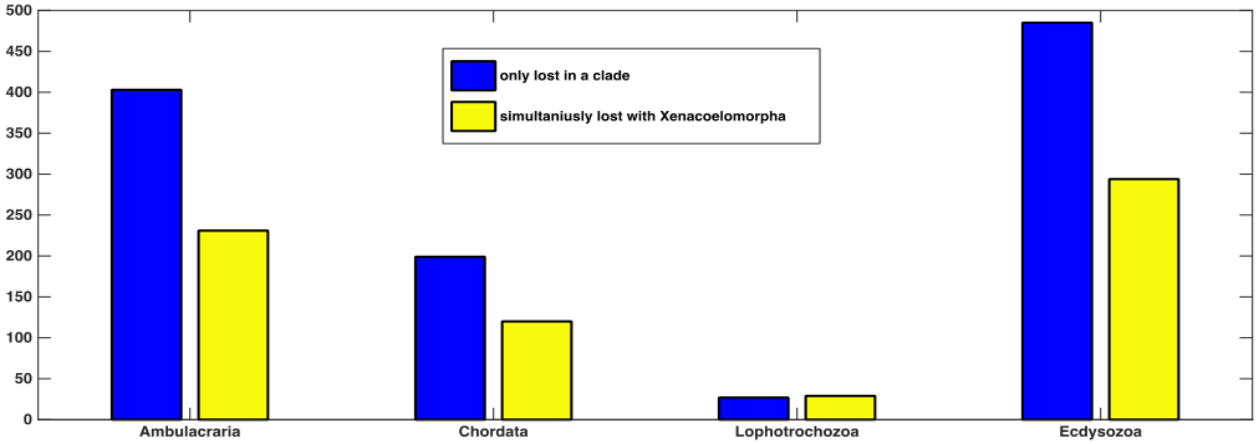


Figure 5.15 Xenacoelomorpha lost more genes simultaneously with Ambulacraria and Ecdysozoa then Chordata and Lophotrochozoa.The number of simultaneous gene losses of the 30,176ancestral gene families shown in yellow. The number of gene family losses specific to a clade shown in blue.

5.3.5.2 Apparent similarity of ancestral Xenacoelomorpha gene content with Xenambulacraria.

Next, we inferred the gene evolutionary event on every branch of 67 species animal phylogeny for the scenario A and B, according to the same protocol as described in Section 5.3.4. For the scenario B) (see Figure 5.17) we inferred 16,443 genes at the base of Metazoa, which increased up to 34,541 in the ancestor of Bilateria. The ancestor of Xenacoelomorpha is inferred to have 27,632 genes, 59% of them were inherited from the ancestor of Bilateria(see Figure 5.16). Surprisingly, 15,606 genes were lost in Xenacoelomorpha from the ancestor of Bilateria. We inferred 6,912 gene duplications and 4,204 *de novo* gene creations on a branch leading to Xenacoelomorpha. In contrast, in the scenario A) we inferred larger gene content of 32,645 genes at the ancestor of Xenacoelomorpha, where more 80% of them were identical to Xenambulacraria last common ancestor (see Figure 5.18). Only 9,105 were

inferred to be lost in Xenacoelomorpha from the Xenambulacraria Last Common Ancestor (see Figure 5.16). 3,776 genes came from duplications and 2,817 from *de novo* gene creations on that branch. Consistently with the results with dataset_2 (34 species) the content of ancestral Xenacoelomorpha genome is inferred to be more similar to Xenambulacraria last common ancestor, then to Bilateria last common ancestor. Additionally, we repeated the evolutionary events inference for dataset_0 with different scenarios for the evolution of insects and mammals, and their inferred gene content was always most similar to the ancestor from which they evolved (not shown). This result supports the phylogenetic position of Xenambulacraria as a sister group of Ambulacraria, however only two competing phylogenetic positions were tested here. Because the family inference of clade specific genes is dependent of the leading phylogeny, proper molecular phylogenetic analysis is required to decide which of the scenarios is more likely. Also, surprisingly large number of gene copies were inferred at the base of Bilateria (34,541) and Xenacoelomorpha (32,645), considering the fact that only 12,611 *Xenoturbella* and 20,409 human gene families were grouped into families, suggesting that the lack of accountant for multiple gene losses of the same gene over the course of evolution may influence our analysis. Moreover, the quality of the data influences the result of events reconstruction, and additional duration of each gene family may be necessary to conclusively say if the losses or duplications we observed in the dataset are not caused by the inaccurate gene predictions and missing data.

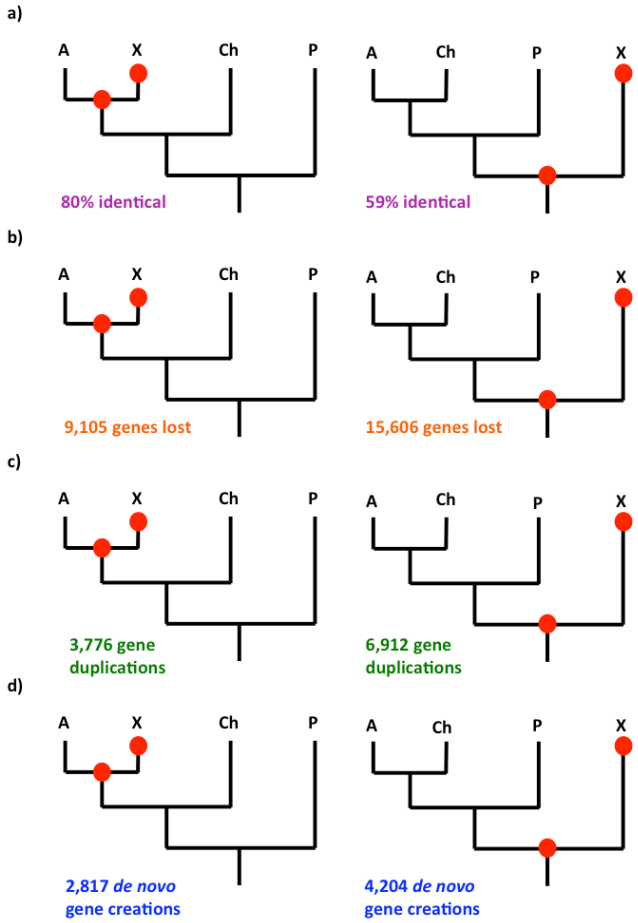


Figure 5.16 Xenacoelomorpha ancestor is more similar to Xenambulacraria ancestor then to Bilateria ancestor. The comparison of ancestral gene content between Xenacoelomorpha ancestor and its presumed predecessor. a) Xenacoelomorpha have more similar gene content to the ancestor of Xenambulacraria ancestor then to Bilateria ancestor(magenta). b) More genes were lost on a branch leading to Xenacoelomorpha in a scenario were Xenacoelomorpha are basal bilateria (orange). c) More genes were duplicated from the ancestor of Bilateria then presumed ancestor of Xenambulacraria. d) More new genes appeared on a branch leading to Xenacoelomorpha in a scenario were Xenacoelomorpha are basal Bilateria.

We found 245,524 OMA orthology groups in the dataset 3 of 67 species. We further used these groups in Chapter 6 for molecular phylogenetic analysis. We propagated GO annotations among the members of a clique within Metazoa according to Gene Ontology propagation in the OMA pipeline as described in Altenhoff et al 2014. Thanks to the repository of orthology groups we were able to produce 1,834,000 new GO annotations. This method is characterized by high precision for general GO terms, and the data can be a good starting point for further gene function analysis.

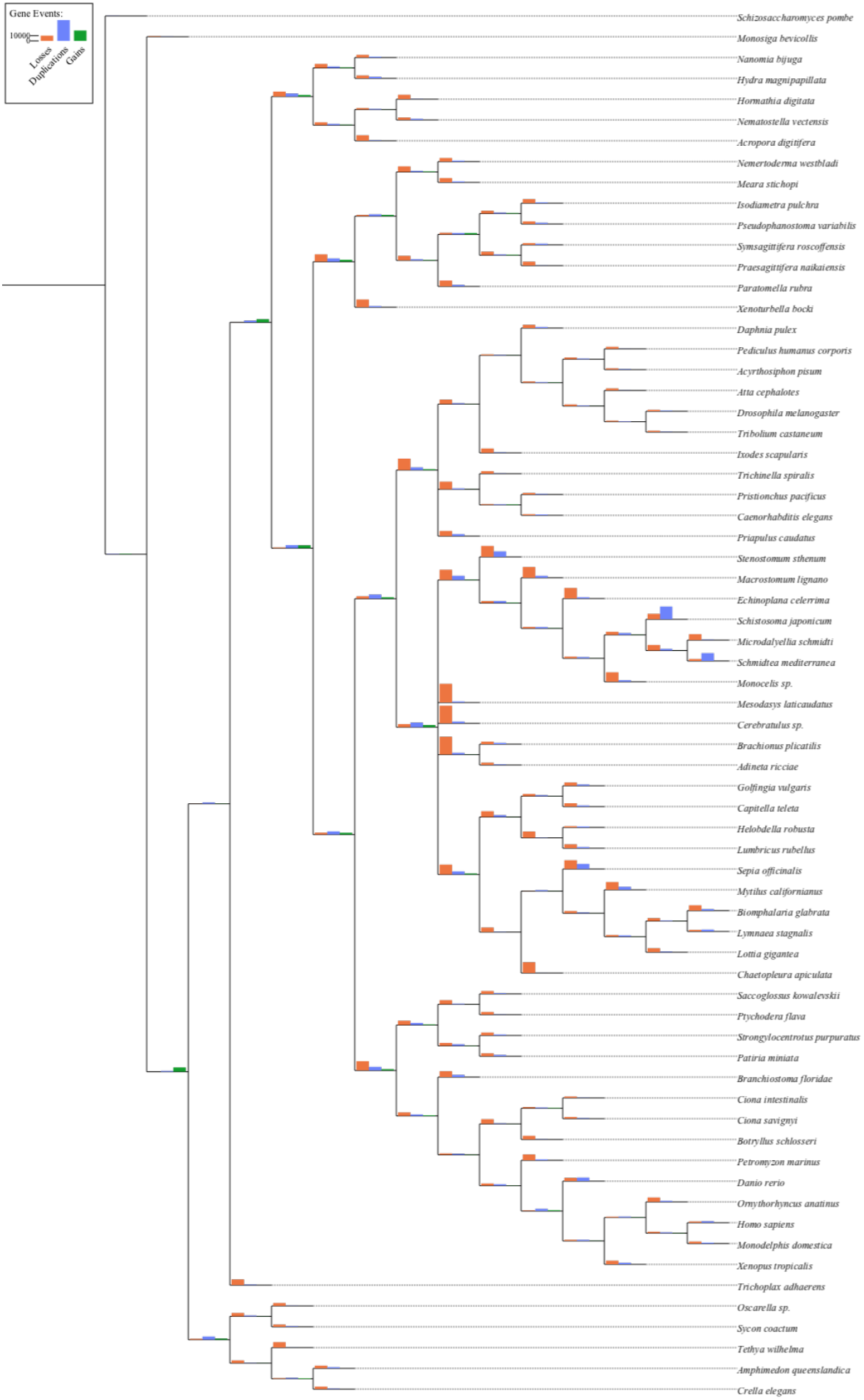
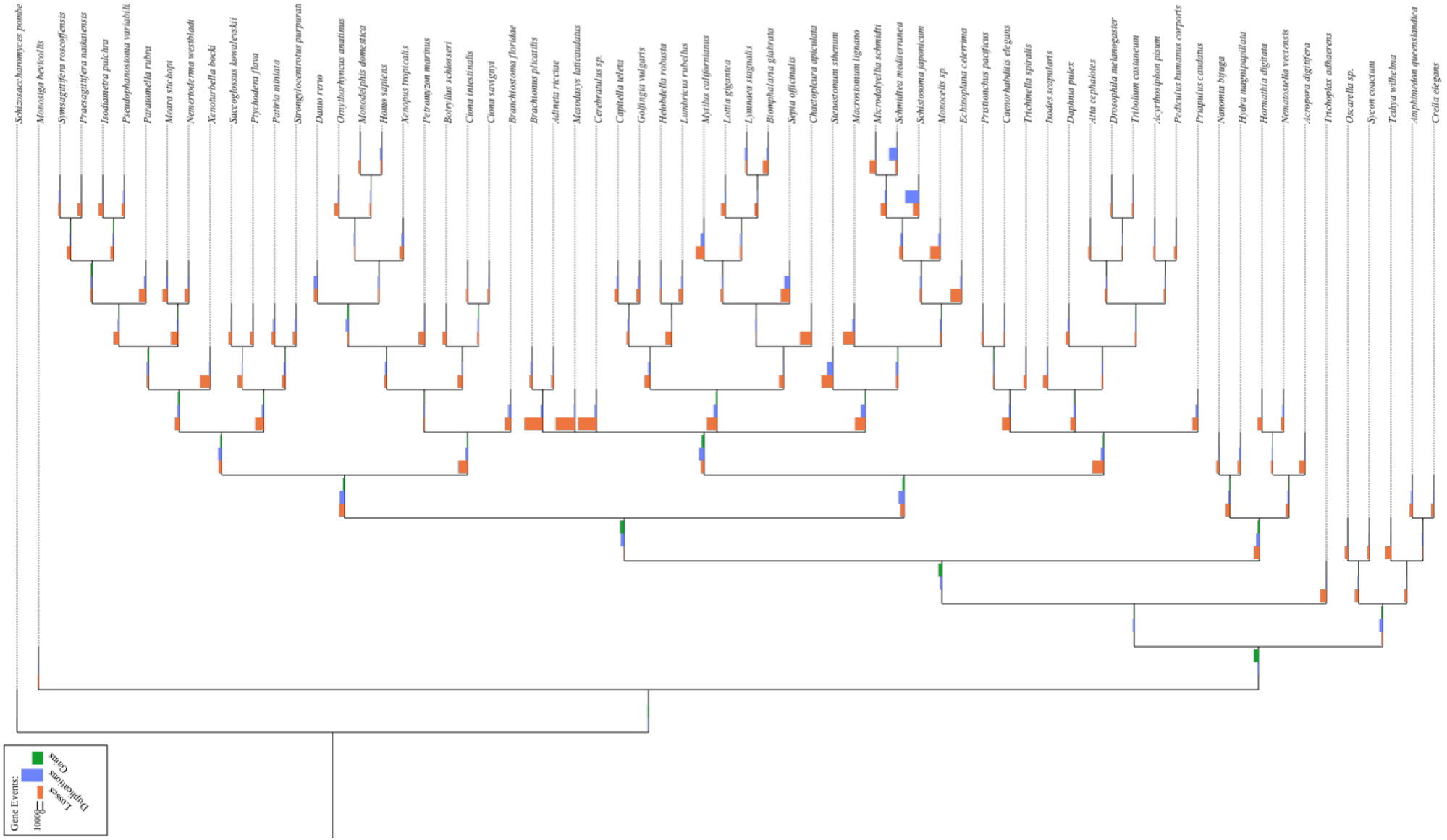


Figure 5.17 Gene family evolution across Metazoa as inferred from 67 proteomes, Xenacoelomorpha are basal Bilateria (scenario B)). Following every single gene copy from its origin until now gene evolutionary events are quantitatively represented on the tree of life. Bars represent an absolute contribution of duplications (blue), losses (red) and *de novo* gene creation (green).

Figure 5.18 Gene family evolution across Metazoa as inferred from 67 proteomes, Xenacoelomorpha are sister group to Ambulacraria (scenario A)). Following every single gene copy from its origin until now gene evolutionary events are quantitatively represented on the tree of life. Bars represent an absolute contribution of duplications (blue), losses (red) and *de novo* gene creation (green).



5.4 Conclusions

We were able to construct large database of Metazoa gene families from 67 species. Database presented here allowed us to follow gene evolutionary events across Metazoa. However, we show that the number of clade specific gene families inferred with OMA standalone is dependent from leading phylogeny, and the proper molecular phylogenetic analysis of the Xenacoelomorpha position remains pivotal for finding clade specific genes in these taxa with OMA standalone. Nevertheless, we reconstructed gene content of ancestral animal proteomes on different levels of animal phylogeny, and follow gene evolutionary events within animal kingdom with 3 scenarios for the evolution of Xenacoelomorpha. We showed that the ancestral gene content of Bilateria expanded rapidly through gene duplication events and *de novo* gene creations from Metazoa ancestor. We showed that little duplication events happen on the branch leading to Xenacoelomorpha then to other clads. Additionally, we showed that the inferred gene content of Xenacoelomorpha in a scenario where Xenacoelomorpha are sister group to Ambulacraria is more similar to Xenambulacraria, then inferred gene content of Xenacoelomorpha to the ancestor of all Bilateria. This result is suggestive for the phylogenetic position of Xenacoelomorpha as a sister group to Ambulacraria. However, missing data, inaccurate gene predictions and fragmented genes influence the inference of the evolutionary events across Metazoan. Further improvement of data quality and more reliable animal phylogeny would improve the reconstruction of gene evolutionary events, and would allow us to more reliably follow the evolution of genes from their birth to death. Animal phylogeny reconstruction from OMA orthology groups presented in Chapter 6 is a step towards better understanding gene evolution within Bilateria.

Chapter 6

Phylogenetic analysis of Xenacoelomorpha based on orthology groups created using whole genomic sequences from 67 Metazoa species.

Final Results and Conclusions.

6. 1 Introduction

The main subject of this work was to find a phylogenetic position of Xenacoelomorpha. To reach this goal I combined datasets and methods presented in Chapters 2-5 to perform large-scale phylogenetic analysis. Moreover, despite the fact that our proteomic dataset contained a high proportion of fragmented gene predictions, for each proteome I was able to identify the presence of core Eukaryotic genes at the similar levels as reported for other reference metazoan proteomes (Chapter 2). Moreover, I was able to identify over 95% of the genes, which were present in 50% of eukaryotes (core and medium abandoned genes), in at least single Xenacoelomorpha proteome, indicating that the new dataset is appropriate for further phylogenetic studies. Additionally, I discovered human and *Burkholderia gladioli* contamination, which I addressed in this Chapter by cleaning the orthology groups based on identical matches froma different species in NCBI database (in cooperation with Hervé Philippe). Motivated by limited number of metazoan species available for analysis with PhylomeDB database, I intended to create the database of orthologous genes, which includes Xenacoelomorpha and many more available sequenced metazoans.

Before constructing such database, I aimed to evaluate available methods for genome scale orthology inference and its application in phylogeny reconstruction using a test dataset. In Chapter four, to choose the best method for the reconstruction of Metazoa phylogenetic dataset, I tested the performance of the three different orthology assignment methods (OMA standalone, CEGMA and OrthoMCL) in phylogenetic multigene alignment construction (superalignment), and phylogeny reconstruction using lophotrochozoan genomic data as a test dataset. OMA standalone produced the largest and most dense phylogenetic superalignment, containing more genes that reconcile the monophyly of Lophotrochozoa than the two other methods. I also found, that the Lophotrochozoa tree reconstructed with OMA standalone was mostly consistent with current literature, by performing phylogenetic analysis using Bayesian and Maximum Likelihood methods (Smith et al. 2011; Kocot et al. 2011). Based on these tests, I concluded that the OMA standalone method is suitable for reconstructing Metazoa phylogeny and thus for finding the phylogenetic position of Xenacoelomorpha. Guided by these results, I used OMA standalone to build the database of metazoan orthologous groups (OMA orthology groups) and gene families (genes that share the common ancestral gene (OMA hierarchical groups) based on proteomic data from 67 species (58 Metazoans including 52 Bilaterians). The analysis of ancestral animal gene content inference (inferred from gene families) shows that Xenacoelomorpha ancestor gene content is more similar to presumed Xenambulacraria last common ancestor then to Bilateria last common ancestor. Additionally, we inferred least gene duplications, losses and gains on a branch leading to Xenacoelomorpha when assuming the scenario where Xenacoelomorpha are sister to Ambulacraria. Therefore, I prefer that phylogenetic positions more then the phylogenetic position of Xenacoelomorpha at the base of Bilateria or at the base of deuterostomes.

The phylogenetic position of Xenacoelomorpha is crucial for understanding the evolution of early Bilateria. As, the phylogenetic position of Xenacoelomorpha is debated in the literature (Lariot et al.2008; Hejnol et al. 2009; Philippe et al. 2011; Cannon et al. 2016; Rouse et al. 2016), we aimed to improve on the previous molecular phylogenetic analysis. To do that we gathered the whole proteomic information from 67 Metazoa. As we established in Chapter 4 that OMA standalone

is the most suitable method for construction of large phylogenetic matrices, we used OMA standalone pipeline described in Chapter 4 together with the large dataset of proteomes described in Chapter 5. For our analysis we choose the Bayesian method of phylogeny inference together with the CAT+GTR+G4 site heterogeneous model, as it was shown previously that this models is the least susceptible to LBA artifact (Lartillot et al. 2004; Philippe et al. 2004). We also observed in Chapter 4, during molecular phylogeny the inference of Lophorthochozoa that Bayesian phylogeny reconstruction with CAT+GTR+G4 site heterogeneous model is more consistent with the literature than Maximum Likelihood reconstruction with GTR+G4 model. The CAT+GTR+G4 site heterogeneous model assumes the heterogeneity of the substitution process across sites and over time, and in addition to the well known site-specific rates (G4), have site-specific equilibrium frequency profiles (CAT), meaning that substitution process at any given site is, on average, confined to a very restricted set of amino acids (i.e. for hydrophobic positions). This allows the modeling of the positions with certain physicochemical properties in a way that the substitutions occur more frequently between amino acids with such property (i.e. for hydrophobic amino acids), and fits the data better. The results of phylogeny reconstruction based on the most informative orthology groups inferred from 67 Metazoa proteomes are presented here.

6.2 Methods

OMA orthologous groups in which more than 50% of species were represented were selected for further analysis (see Figure 4.2). Protein sequences from each orthology group containing sequences from at least 64 species were aligned using MUSCLE (Edgar et al. 2004), with default settings. Additional sequences from Porifera were added to the dataset and cleaning the orthology groups based on identical matches from different species in NCBI database was performed by Hervé Philippe. Unreliable portions of the alignment were removed from the alignments using trimAl (Capella-Gutierrez et al. 2009), with default settings. The final alignment was created by concatenation of all alignments from 2,162 OMA orthology groups with 15 or more genes (>=50% complete, similarly we concatenated 438 CEGMA core orthology groups and 484

OrthoMCL groups with 15 or more members). Missing sequences were represented by gaps. The full alignment was finally reduced to sites with more then 70% occupancy. Bayesian inference was conducted with PhyloBayes (PhyloBayes version1.5a in open mpi version 1.8.1 environment on a UCL Computer Science Cluster) using the CAT GTR model, in parallel using 32 CPU cores per chain (Lartillot et al. 2013). Two independent MCMC chains of 1,000 generations each were run on each alignment. The first 100 trees (10%) were discarded as burn-in for each MCMC run prior to convergence (i.e., when maximum discrepancies across chains <0.3).

6.3 Results

I managed to assemble coding sequence data from 67 animal species, including 56 species of Metazoa. I identified 245,524 orthology groups using OMA standalone software, and produced over 1,500,000 new functional predictions (GO terms annotations based on propagating previously annotated function from experimental and electronic evidence). Within these groups, I identified a subset of 3,683 orthology groups, which contained genes present in at least 50% (34) of analysed species and in at least one member of Xenacoelomorpha. These genes in these orthology groups were aligned, trimmed and concatenated into a superalignment. In cooperation with Hervé Philippe, the contaminations were removed from the superalignment. This done with two procedures. First, the sequences with identical matches from a different species in NCBI database were removed. Second, sequences with bizarrely long branch lengths on a gene tree were removed (supposedly psudogenes). Moreover, additional sequences from species without sequence genomes were added (from Ambulacraria, Porifera, Cnidaria and the acoel *Hofstenia miamia*).

Finally, we produced a large (>350,000 positions) and taxonomically broad phylogenomic dataset (9 Xenacoelomorpha species, 8 Chordata, 15 Ambulacraria, and 13 Protostomia). The dataset contains very few missing data (minimum 60% complete per position with data present within the total alignment), with an average 76% complete (as the percentage of positions with data present within the total alignment), and all but 6 taxa (*Petromyzon marinus, Chondrilla nucula, Meara stichopi, Ircinia fasciculata, Schizocardium brasiliense, Cephalodiscus gracilis*) with more

then 56% complete. We performed Bayesian tree reconstruction using site-heterogeneous model (PhyloBayes CAT+GTR+G4 model), which has site-specific equilibrium frequency profiles, to reconstruct the phylogeny of Metazoa.

The Bayesian analysis supports the monophyly of Xenacoelomorpha (Xenoturbellida + Acoelomorpha) (see Figure 6.1), and Xenacoelomorpha were found to be sister group of Ambulacraria. However, two Markov chains in the Bayesian analysis are not convergent, and support different position of Xenacoelomorpha (1st chain Xenacoelomorpha were found to be sister group of Ambulacraria; 2nd chain Xenacoelomorpha as basal Bilateria), this discrepancy is also reflected by low posterior probability in the consensus tree. The phylogeny recovers the well-established topology within five major Bilateria clades (Chordata, Ambulacraria, Xenacoelomorpha, Lophotrochozoa and Ecdysozoa), and supports the basal position of Porifera within Metazoa (Philippe et al. 2009). Surprisingly, the phylogeny does not support the monophyly of Deuterostomia, and places Chordata as a sister group to Protostomia.

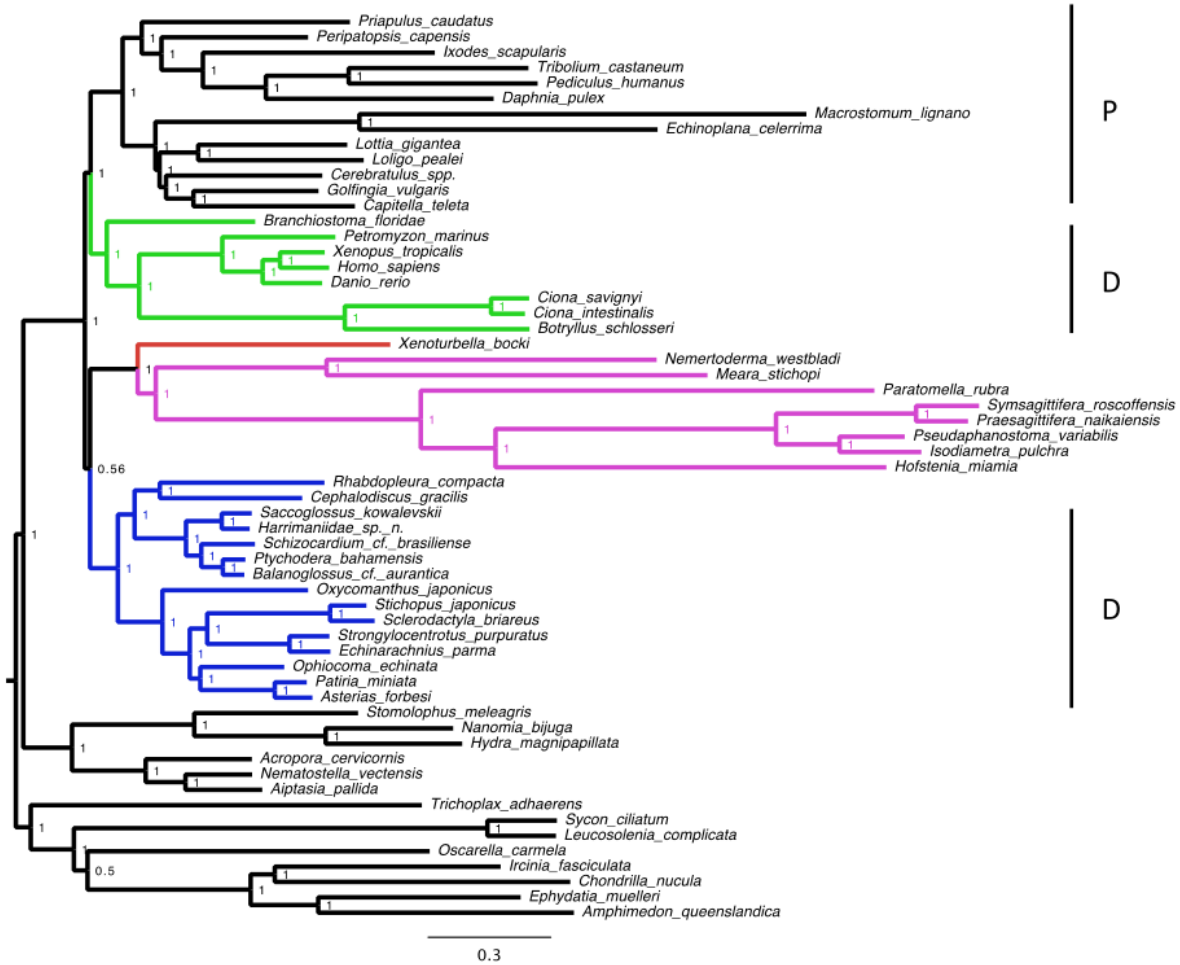


Figure 6.1. Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 model from full 350,090 Amino Acid positions alignment. The phylogeny supports the monophyly of Xenacoelomorpha (Xenoturbellida (red) and Acoelomorpha (magenta)), and places Xenacoelomorpha as a sister group to Ambulacraria (blue) with weak support (PP=0.56). The phylogeny does not support the monophyly of Deuterostomia (D) and supports Chordata (green) as a sister group to Protostomia (black; P). MaxDiff = 1.0; MeanDiff = 0.0163313.

To get more confidence in the relative position of 4 major clades within Bilateria (Chordata, Ambulacraria, Xenacoelomorpha and Protostomia), we have performed a jackknife resampling (The analysis in which generate a simple random sample without replacement from the superalignment). We have subsampled the superalignment 100 times, by selecting 20,000 positions at random from the full data set (generated by Max Telford). Each dataset was analysed for 300 cycles using CAT+GTR+G4 model in PhyloBayes. For each of 100 Markov chains the consensus trees were inferred from last 100 cycles, of each Bayesian inference. The jackknife tree is a consensus from the trees calculated for each sample (see Figure 6.2). The jackknife tree supports the monophyly of Xenacoelomorpha (Xenoturbellida + Acoelomorpha) (see Figure 6.2), but does not support the position of Xenacoelomorpha as a sister group of Ambulacraria. This tree reconciles the same topology within Deuterostomia, Ambulacraria and Chordata as a full tree (see Figure 6.1). Yet, it does not support the position of *Cerebratulus spp.* and Platyhelminthes (*Macrostomum lignano, Echinoplana celerina*) within Lophotrochozoa, which results in polytomy. Similar to the Bayesian analysis (see Figure 6.1), the jackknife tree does not support the monophyly of deuterostomes (indicated on a tree with D"). The relative position of Protostomia, Deuterostomia, Chordata and Xenacoelomorpha is unresolved, which I indicated by polytomy on the jackknife tree (see Figure 6.2).

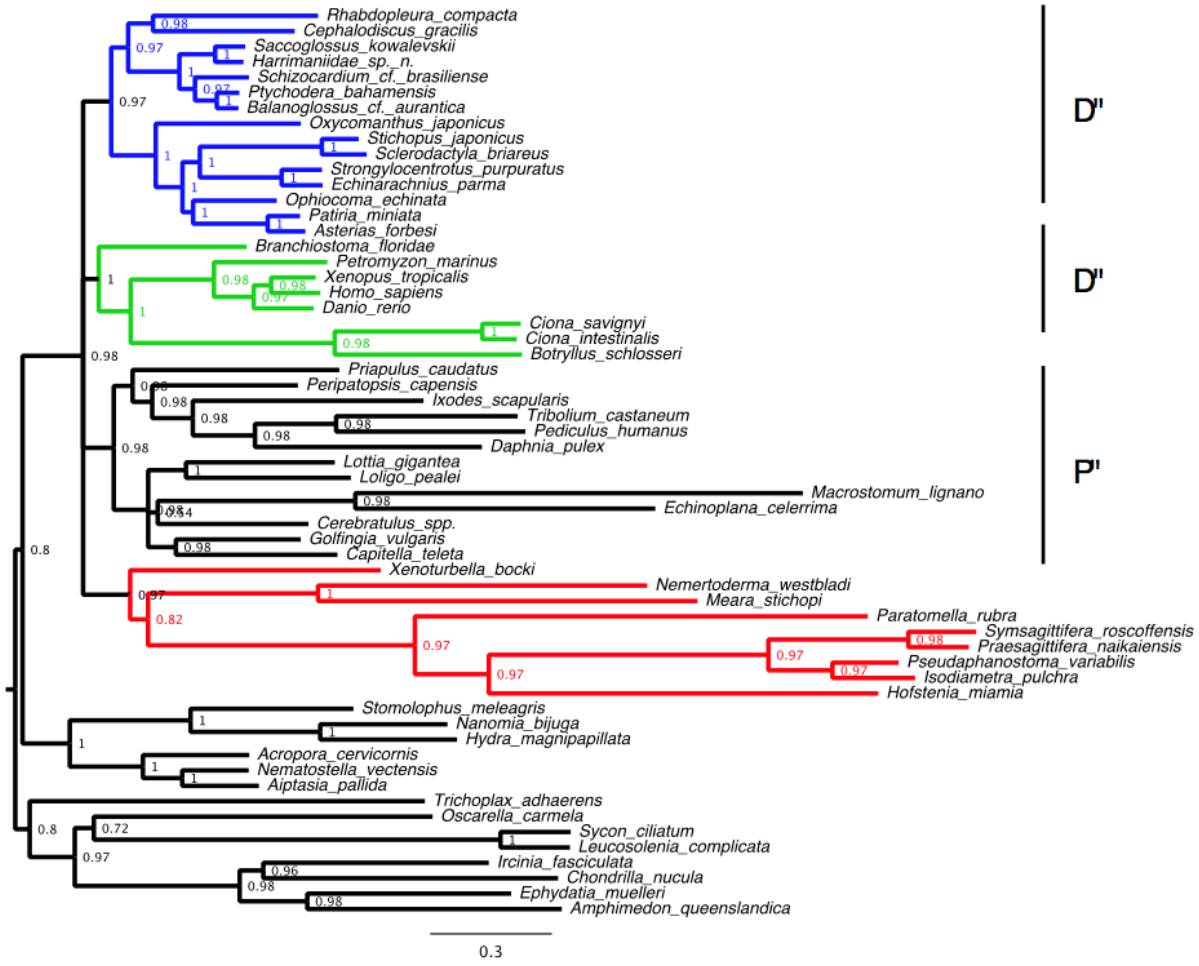


Figure 6.2 The jackknife analysis of 100 datasets of 20,000 amino acids each, inferred using CAT+GTR+G4 model in PhyloBayes. Values at nodes indicate proportion of replicates in which the node is found (1 corresponds to 100% jackknife). The topology is largely the same as the full analysis shown in Figure 6.1, and most clades receive high support. There is no clear support for the monophyly of Deuterostomia indicated by the polytomy at the base of the Bilateria (polyphyletic deuterostomes marked as D", protostomes marked as P"). Scale bar indicates number of substitutions per site.

Because Xenacoelomorpha are fast-evolving they may be grouped with the second fast evolving linage (Porifera, or Cnidaria) or other distant out-group, as the result of long branch attraction (LBA). Even though we have used the site heterogeneous mixed model, which is shown to minimize the effects of LBA (Lartillot et al. 2008), the artefact may still interfere with the result of the analysis, due to low resolution in inferring of the ancestral nodes. Considering that the close relation of Acoelomorpha and Xenoturbellida have already been well established (Franzen and Afzelius 1987; Lundin 1998, 2001; Raikova et al. 2000; Rohde et al. 1988), to eliminate the influence of fast evolving linages on the tree topology inference we removed fast

evolving Acoelomorpha from the dataset. Thus, in a new dataset the relatively short branched *Xenoturbella bocki* was the only representative of Xenacoelomorpha.

We have performed Bayesian tree reconstruction analysis in PhyloBayes using CAT+GTR+G4 model on the reduced dataset (see Figure 6.3). The phylogeny supports Xenoturbella as a sister group of Ambulacraria with high posterior probability. Both Markov chains from Bayesian analysis were convergent (passed the maxdiff: 0.209091 < 0.3; meandiff: 0.00211203 < 0.1 test). The tree is better supported with all ancestral nodes having posterior probability equal to 1. Similar to the full tree (see Figure 6.1), the Bayesian analysis with the reduced dataset does not support the monophyly of deuterostomes, and places Chordata as a sister group of Protostomia (see Figure 6.3).

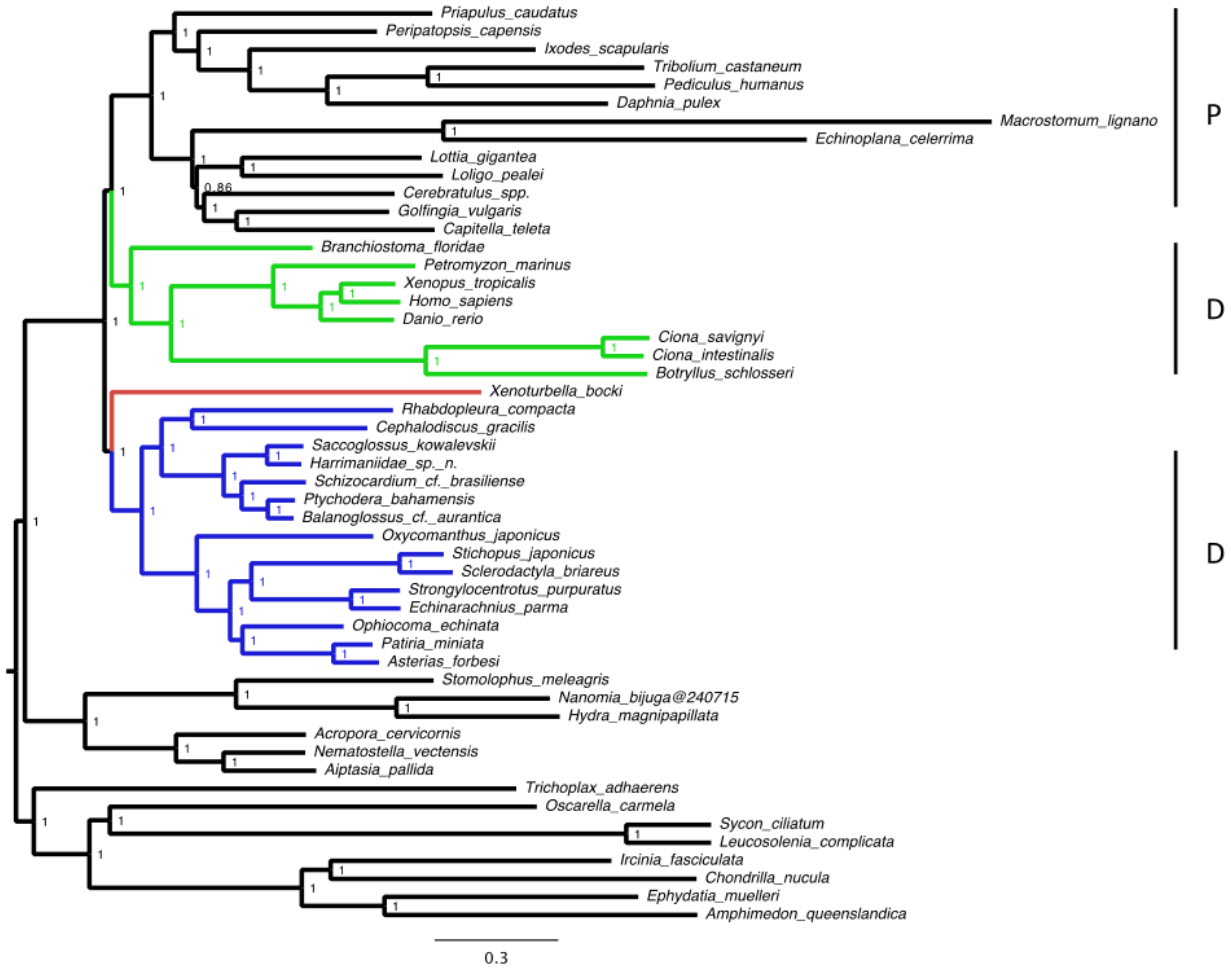


Figure 6.3 Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 model based on the full 350,090 amino acid positions alignment, without Acoelomorpha. There is support for a sister group relationship between Chordata (green) and Protostomia (black), as well as Xenoturbellida (red) and Ambulacraria (blue). Polyphyletic deuterostomes marked as D, protostomes marked as P. Maxdiff: 0.209091; meandiff : 0.00211203. Scale bar indicates number of substitutions per site.

To get more support for the inferred Bayesian phylogeny we have performed jackknife resampling analysis (20,000 positions at random from the full data set; samples prepared by my supervisor Max Telford) in PhyloBayes using site heterogeneous CAT+GTR+G4 model for 300 cycles (see Figure 6.4). The topology is largely the same as in the analysis of full alignment of the reduced dataset (see Figure 6.3). 70% of the sampled jackknife trees support the position of Xenoturbellida as a sister group to Ambulacraria. 60% of the sampled jackknife trees support Chordata as a sister group to Protostomia. It is important to note, that our analysis is still in progress, the jackknife replicates were run only for 300 cycles, and we did not yet investigate the convergence of each chain. We plan to perform longer jackknife analysis to further investigate the support for the consensus topology.

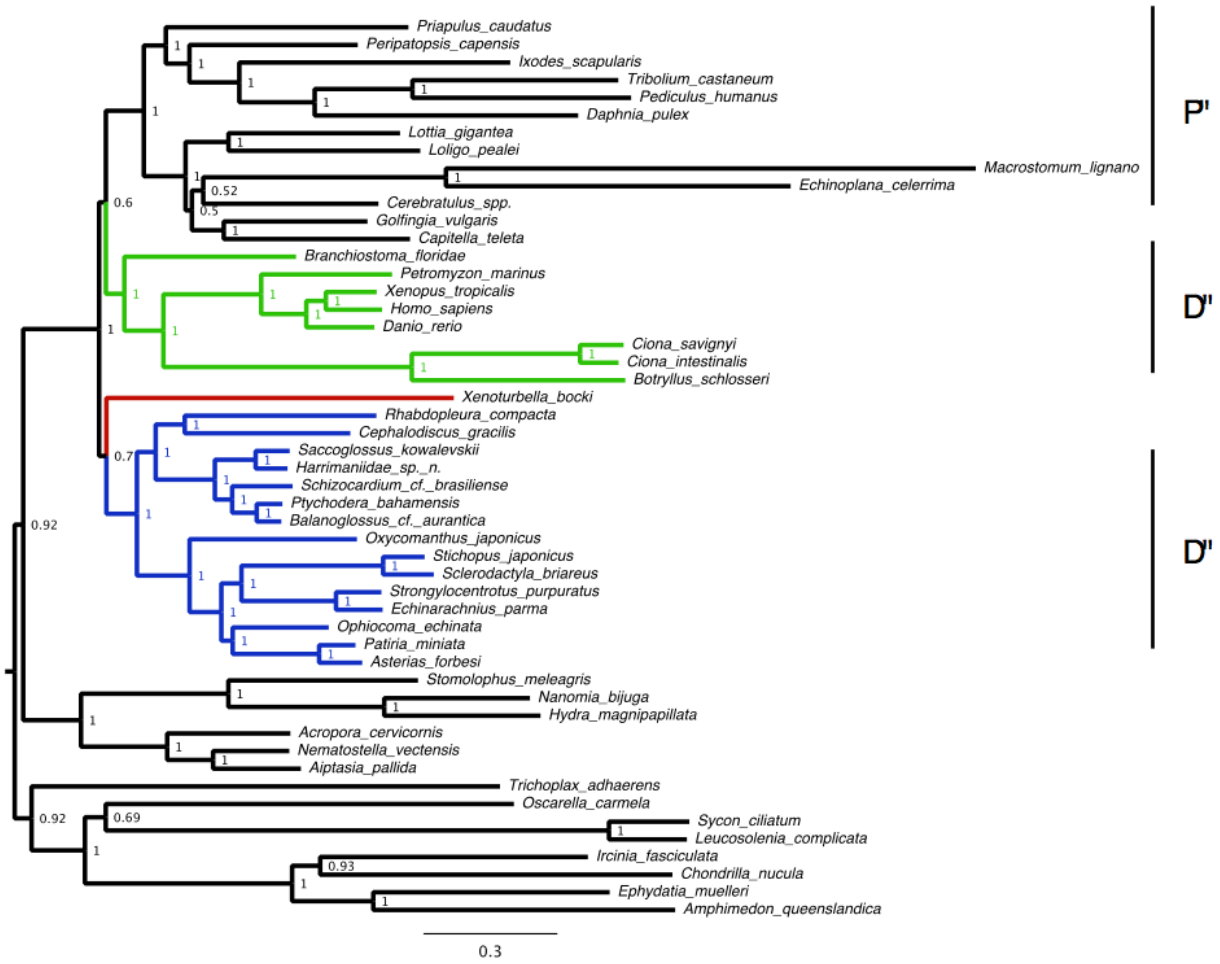


Figure 6.4 The jackknife analysis of 100 subsets of 20,000 amino acids each, produced using the PhyloBayes CAT+GTR+G4 model, without Acoelomorpha. Values at nodes indicate proportion of replicates in which the node is found (1 corresponds to 100% jackknife). The topology is largely the same as the full analysis shown in Figure 6.3, and most clades receive high support. 70% of the jackknife samples supports Xenoturbellida (red) as a sister group to Ambulacraria (blue; D). 60% of the jackknife samples supports Chordata (green; D) as a sister group to Protostomia (black; P). Scale bar indicates number of substitutions per site.

Next, we have considered that genes, which are sorted in the same way as the animal speciation events, will reconstruct the animal evolutionary history better than whose which not. Those genes should reconstruct the monophyly of known monophyletic groups. Therefore, we have ranked the orthology groups, based on their ability to recover the monophyly of known monophyletic groups. First, we filtered our full dataset (see Figure 6.1) for the phylogenetic signal, by keeping only orthology groups, for which gene trees reconsolidate the monophyly of Ambulacraria (the ranking prepared by Max Telford). We have chosen a subsample of 31,000 positions to perform a Bayesian analysis with site heterogeneous CAT+GTR+G4 model in PhyloBayes.

The Bayesian tree calculated based on Ambulacraria monophyletic genes supports the position of Xenacoelomorpha as a sister group of Ambulacraria with posterior probability of 0.92, and is consistent with our result, in which we excluded long branching Acoelomorpha (see Figure 6.3, 6.4) and also supports the previous studies by Philippe and others (Bourlat et al. 2006; Telford et al. 2008; Philippe et al. 2009, 2011; Nakando et al. 2013). The tree reconciles the branching patterns within Chordata, Ambulacraria and Protostomia (see Figure 6.1) consistently with previous studies (Hejnol et al. 2009; Philippe et al. 2011), and receive high support. Within Xenacoelomorpha, Nemertodermatida (*Meara Stichopi* and *Nemertoderma westbladi*) are basally positioned to *Xenoturbella* and Acoelomorpha with posterior probability 0.94. This analysis also does not support the monophyly of Deuterostomes, as Chordata are basally branching Bilaterians with high support posterior probability 1.00. The latter result is neither in agreement with full dataset phylogeny (see Figure 6.1), nor with the analysis of reduced dataset, in which we excluded long branching Acoelomorpha (see Figure 6.3). However, the result that Deuterostomes are paraphyletic clade is consistent between these three phylogenies and supports previous findings by Lartillot et al. 2008, but is not in agreement with other phylogenies published so far (Hejnol et al. 2009; Philippe et al. 2011; Cannon et al. 2016).

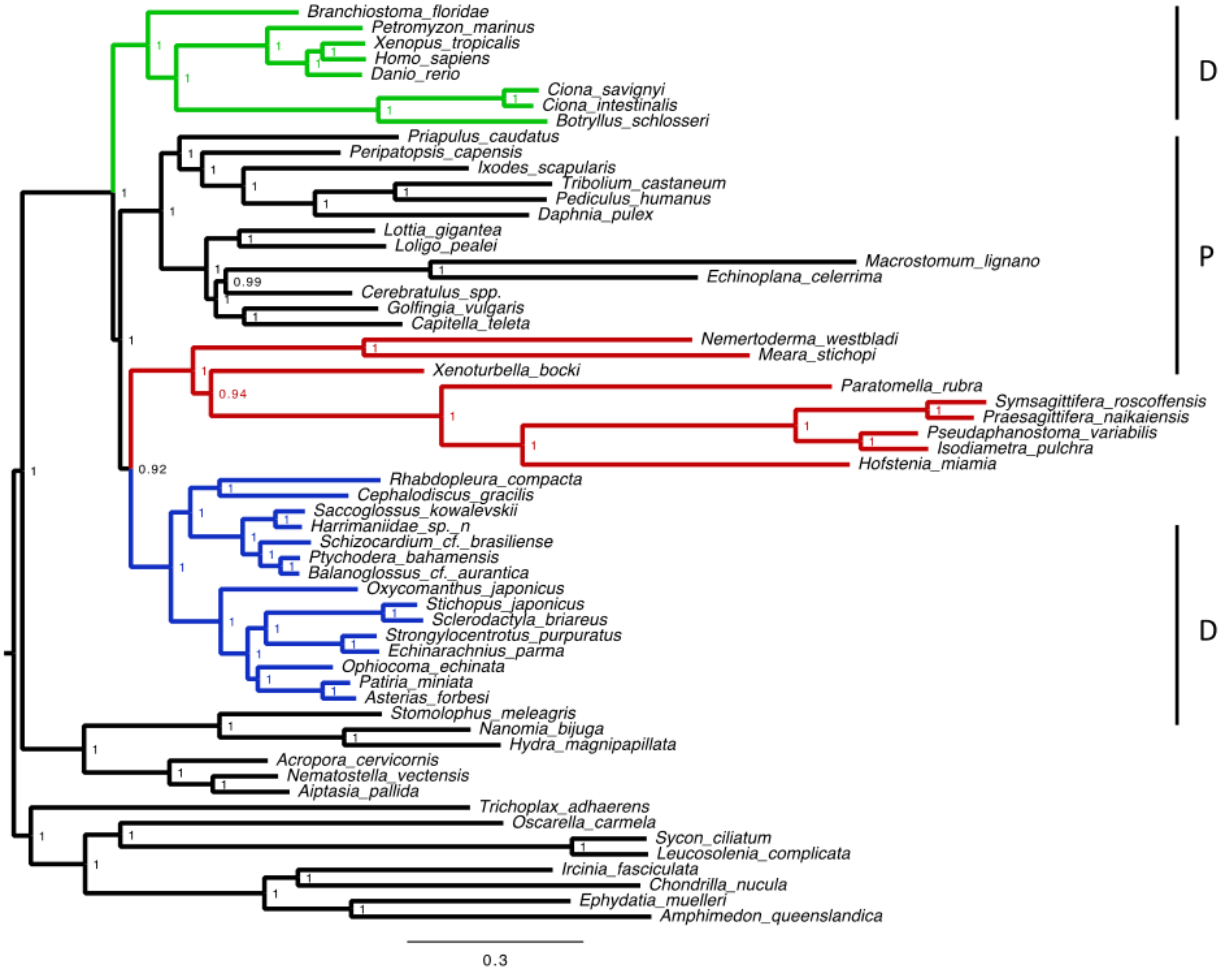


Figure 6.5 Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 on 31,000 amino acid positions superalignment of genes that reconsolidate monophyletic Ambulacraria. There is strong support for a position of Chordata (Green, D) at the base of bilateria. Protostomia (black, P) (pp=1), are sister group to Xenoambulacraria (Xenacoelomorpha (red) and Ambulacraria (blue, D)) with high support pp=0.92. Scale bar indicates number of substitutions per site.

Similarly, we selected orthology groups, which gene trees reconcile the monophyly of Protostomia best, and chose a subsample of 31,000 positions to perform a Bayesian analysis with site heterogeneous CAT+GTR+G4 model in PhyloBayes (the ranking prepared by Max Telford). The phylogeny supports the position of Xenacoelomorpha as a sister group to Acoelomorpha with high posterior probability 1.00 (see Figure 6.6) and recovers the same topology within Xenacoelomorpha and Ambulacraria as the topology obtained with full dataset (see Figure 6.1). The obtained tree topology is consistent with the analysis using Ambulacraria monophyletic genes (see Figure 6.5) and the reduced dataset that excluded long branching Acoelomorpha (see Figure 6.3). The tree topology is also consistent with previous studies (Bourlat et al. 2006; Telford et al. 2008; Philippe et al. 2009, 2011; Nakando et al. 2013).

Finally, the result supports paraphyletic Deuterostomes, as Chordata are sister group to Protostomia (see Figure 6.6).

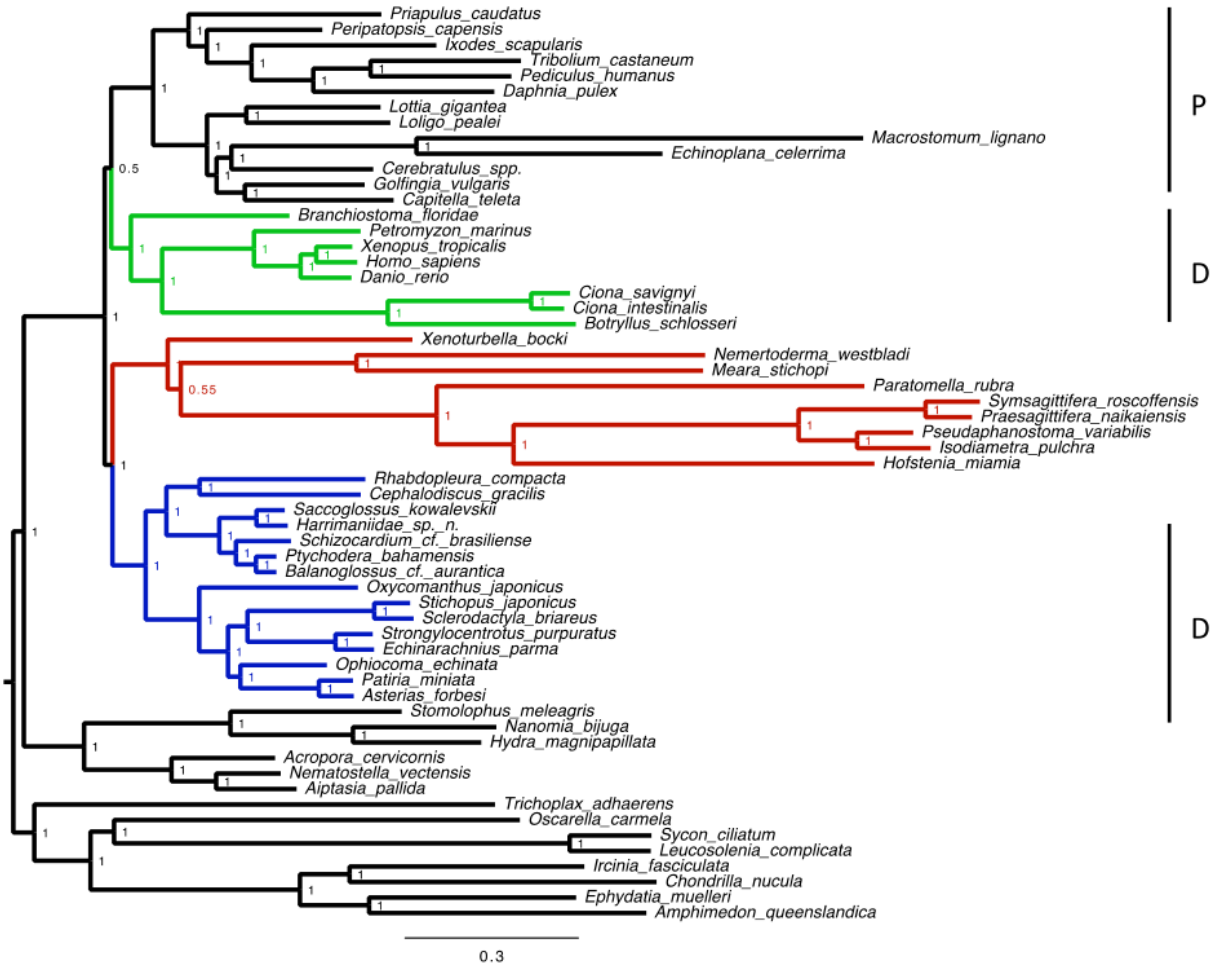


Figure 6.6 Phylogeny inferred using PhyloBayes with the Site-Heterogeneous CAT+GTR+G4 on 31,000 amino acid positions superalignment of genes that reconsolidate monophyletic Protostomia. Xenacoelomorpha (red) are sister group of Ambulacraria (blue, D) with high support pp=1. There is weak support for a position of Chordata (Green, D) as a sister group of Protostomia (black, P) pp=1. Alternative conflicting topology was supporting the position of Chordata (Green, D) at the base of Bilateria. Scale bar indicates number of substitutions per site.

6.4 Conclusions

Several lines of evidence presented in this thesis supports the view that Xenacoelomorpha are sister group to Ambulacraria, and thus are neither basal to Bilateria or Deuterostomia. (i) Deuterostome specific genes were found in the Xenacoelomorpha proteomes. (ii) Xenacoelomorpha and Ambulacraria share common gene losses. (iii) The smallest number of gene loss events was inferred once Xenacoelomorpha was placed on phylogenetic tree as sister to Ambulacraria. (iv) Proteome based phylogeny with the dataset without fast evolving acoels, as well as, the phylogenies with the dataset restricted to genes, which trees supported the monophyly of Ambulacraria and Protostomia, all support the position of Xenacoelomorpha as a sister group to Ambulacraria.

Our support for the position of Xenacoelomorpha as a sister group to Ambulacraria is consistent with large-scale Bayesian analysis of 66 taxa and 38,330 amino acids positions by Philippe (Philippe et al. 2011), but is not consistent with the analysis presented by Hejnol group (Hejnol et al. 2009; Cannon et al. 2016). The analysis from 2009 shows the phylogeny inference of only 19% gene occupied large supermatrix (94 taxa and 270 580 amino acids positions (Hejnol et al. 2009)) only with Maximum Likelihood method. The recently improved analysis by the same group also supports the basal position of Xenacoelomorpha with Maximum Likelihood method, and with Bayesian method with CAT+GTR+G4. However, the Bayesian phylogeny with CAT+GTR+G4 model was only inferred on a small set of 212 genes from of 76 taxa based on the 69% complete phylogenetic supermatrix of 44,896 amino acids (Cannon et al. 2016). We improved on previous analyses by constructing larger and more complete dataset of carefully sampled 67 animal species with maximum diversity, which includes proteomes from 9 Xenacoelomorpha and 15 Ambulacraria species. We used larger collection of 3,683 orthologous genes inferred with best performing orthology inference method (based on the analysis from Chapter 3). From this we have constructed larger superalignment of over 350,000 positions with a better 76% completeness and accounted for the contamination. To account for Long Branch

Attraction and cases of heterogeneity amino acid composition across lineages, we used better-fitted site heterogeneous CAT+GTR+G4 model (Lartillot et al. 2004; Philippe et al. 2004).

If Xenacoelomorpha are sister to Ambulacraria, then their simple morphology is likely a result of secondary simplification from a more complex Xenambulacraria Last Common Ancestor. We should consider, how this simplification has arisen. In that case Xenacoelomorpha secondarly lost anus, endostyle, gill slits, postanal tail, through gut cephalic brain and deuterostomy. Even though similarly high number of gene losses was observed on branches leading to Xenacoelomorpha, Ambulacraria, Nematodes and Platyhelminthes, a low level of gene duplications was observed exclusively for Xenacoelomorpha, and may play a role in their morphological simplification (see Chapter 5). However, as the outcome of gene family analysis was influenced by genomic data quality and the leading phylogeny, more research is necessary to fully understand the phylogenetic position of Xenacoelomorpha and genetic reasons of their morphological simplicity. Nevertheless, the methods described in this thesis are highly applicable and will likely provide more information once improved sequence data and well-supported leading phylogeny will be available.

The unexpected result of our phylogenetic analysis of metazoans is paraphyly of Deuterostomes. The analysis of reduced dataset, lacking the fast evolving Acoelomorpha taxons, and the analysis of dataset consisting of genes, which phylogenies supported the monophyly of Protostomia, both support monophyletic Xenambulacraria, and Chordata as a sister group to Protostomia. While the analysis of dataset consisting of genes, which phylogeny supported monophyly of Ambulacraria, supports basal position of Chordata within Bilateria, which is consistent with results obtained previously by Lartillot (Lartillot et al. 2008). However, other Metazoa phylogenies published so far support monophyletic deuterostomes (Hejnol et al. 2009; Bourlat et al. 2006; Telford 2008; Philippe et al. 2009, 2011; Nakando et al. 2013), with very short branch leading from Urbilateria to Urdeuterostome. One explanation for these conflicting branching patterns could be the lack of phylogenetic resolution at the base of Bilateria due to radiation of Chordata, Protostomia and Xenambulacraria within a short period of time. This, together with the lack of available extant early branching species at the base of Bilateria, makes this branching pattern

difficult to resolve. Other explanation for the previously found monophyly of deuterostomes (Hejnol et al. 2009; Bourlat et al. 2006; Telford et al. 2008; Philippe et al. 2009, 2011; Nakando et al. 2013**)**, is that the fast-evolving protostomes were clustered with Cnidaria due to long branch attraction. Our findings, as well as previous findings **(**Lartillot et al. 2008), using Site-Heterogeneous CAT+GTR+G4 model, which accounts for Long Branch Attraction by calculating substitution rates of the differently evolving sites with different substitution matrixes, do not support the monophyly of Deuterostomes. Paraphyly of deuterostomes, if true, would markedly change our understanding of the evolution of Bilateria, as we might consider that such developmental characteristics as radial cleavage, deuterostomes gastrulation and enterocoelic mode of formation of body cavity considered as typical for deuterostomes, were present already in Urbilateria. Moreover, some of these characteristics can be even found in more basally branching protostomes (Matus et al. 2006; Marlétaz et al 2006**)**. For example deuterostomy (condition in which the blastopore forms the anus of the adult animal), which is present in all deuterostomes, is also present in some protostomes such as brachiopods (members of Lophotrochozoa). Recently Martín-Durán et al. 2012, based on the observation of embryonic development in basal Ecdysozoa; *Priapulus caudatus,* suggested that last common ancestor to the Deuterostomia and Protostomia exhibited deuterostomic development. This suggests that, deuterostomy present in Urbilateria, was inherited by extant chordates and xenambulacrarians but was lost within Lophotrochozoa and Ectodysozoa linages. Similarly radial cleavage, another characteristic presumed to be typical for deuterostomes, is most likely a basal character state of metazoans, as it was found in poriferans, cnidarians and ctenophores, and also in chordates and xenambulacrarians. While the alternative way of embryonic development type, such as spiral cleavage, has a single origin in a sub-group of protostomes that include annelids, molluscs, nemerteans and Platyhelminthes, and is not an ancestor protostome state. Moreover, if Deuterostomia are not monophyletic the lack of typical deuterostomic characteristics, such as gill slits, endostyle or deuterostomy in Xenacoelomorpha could be explained by two independent losses. Both lineages could have lost them during the course of evolution (figure 6.7).

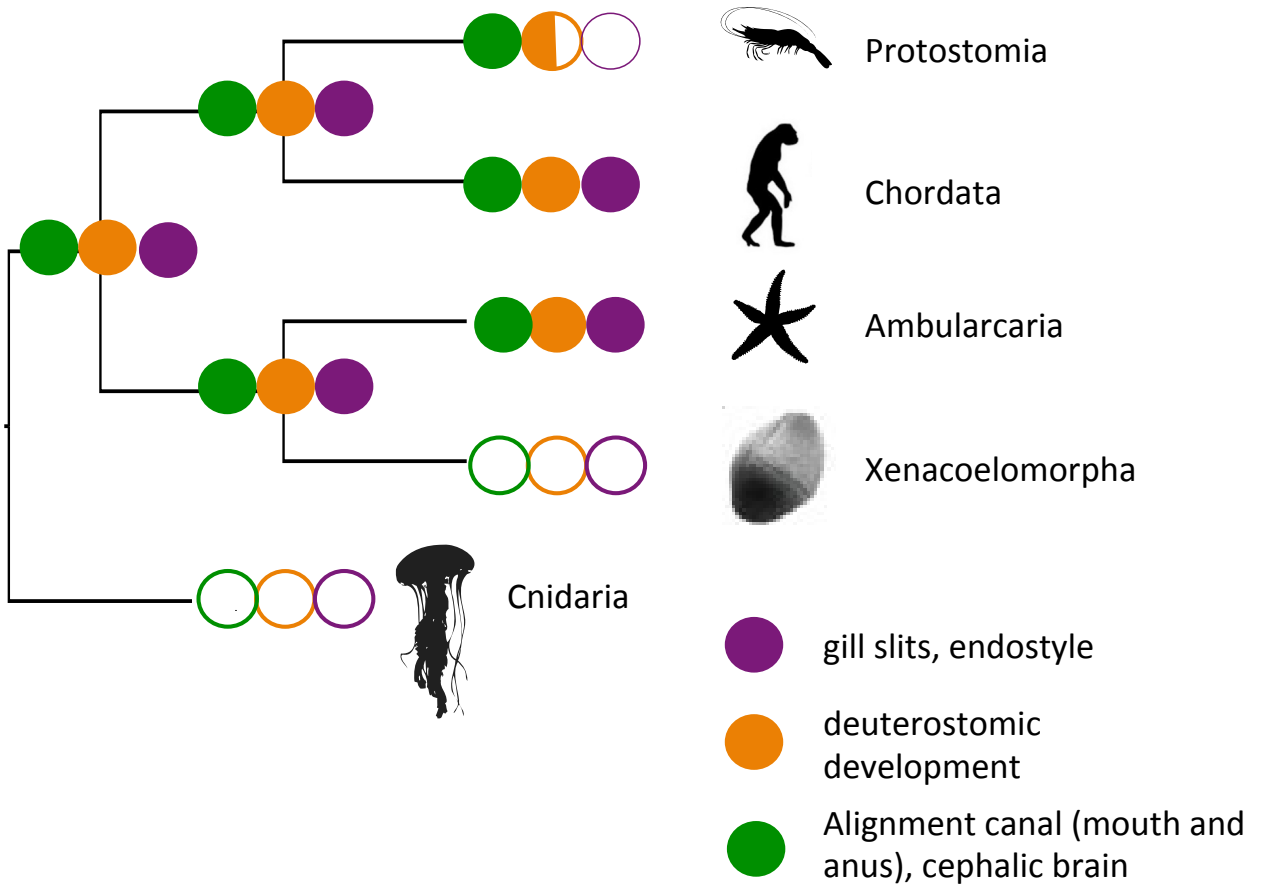


Figure 6.7 Evolution of characteristics within Metazoa according to the scenario, where deuterostomes are not monophyletic (supported by molecular phylogeny with the Site-Heterogeneous CAT+GTR+G4 (Figure 6.1; 6.3; 6.4; 6.6)). The presence of a characteristics is indicated by full circle, lack of a characteristic is indicated by empty circle, while partial presence of a characteristic is indicated by half-pie. Full circles on the nodes of the phylogeny indicate the presence of gill slits, endostyle, deuterostomy, alignment canal and cephalic brain in Urilateria and other common ancestors. The lack of typical deuterostomic characteristics, such as gill slits, endostyle or deuterostomy in Xenacoelomorpha could be explained by two independent losses.

6.4 Outline

Further work is needed to get more confidence in the topology of Bilateria phylogeny. To get more confidence in our analysis, and improve the phylogenetic signal in coming from the phylogenetic matrix, we plan to further filter our dataset by choosing genes, which reconcile monophyly of the all four main clades of Bilateria (Lophotrochozoa, Ecdysozoa, Chordata, Ambulacraria). Additionally, to reduce the effects of Long Branch Attraction, sites saturation and allow the better fit of evolutionary model we plan to dissect our dataset into classes with increasing rates of evolution and calculate phylogeny for each of these classes (Egger at al. 2015). For each of

the classes we would recalculate the phylogeny the best-fitted model and with both Bayesian and Maximum likelihood methods of tree reconstruction. Further improvement of the data quality will certainly provide more confidence in both gene family and phylogenetic analysis. One step towards that goal would be an improvement of genome quality. Second, would be a additional curating of gene families with the sequence data from other sources and more thorough analysis of ancestral genome content.

References

1. Abildgaard, P. C. (1806). Zoologica danica seu animalium Danicae et Norwegiae rariorum ac mines notorum descriptorum et historia. In O. F. Müller (Ed.), Zoologica Danica 4 (p. 26). Copenhagen (Havnia): N. Mölleri.
2. Achatz, Johannes G., et al. "The Acoela: on their kind and kinships, especially with nemertodermatids and xenoturbellids (Bilateria incertae sedis)." Organisms Diversity & Evolution 13.2 (2013): 267-286.
3. Achatz, Johannes Georg, et al. "Systematic revision of acoels with 9+ 0 sperm ultrastructure (Convolutida) and the influence of sexual conflict on morphology." Journal of Zoological Systematics and Evolutionary Research 48.1 (2010): 9-32.
4. Akaike, Hirotugu. "A new look at the statistical model identification." IEEE transactions on automatic control 19.6 (1974): 716-723.
5. Akaike, Htrotugu. "Maximum likelihood identification of Gaussian autoregressive moving average models." Biometrika 60.2 (1973): 255-265.
6. Altenhoff, Adrian M., and Christophe Dessimoz. "Inferring orthology and paralogy." Evolutionary genomics. Humana Press, 2012. 259-279.
7. Altenhoff, Adrian M., et al. "Inferring hierarchical orthologous groups from orthologous gene pairs." PLoS One 8.1 (2013): e53786.
8. Altenhoff, Adrian M., et al. "The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements." Nucleic acids research (2014): gku1158.
9. Altenhoff, Adrian M., et al. "The OMA orthology database in 2015: function predictions, better plant support, synteny view and other improvements."Nucleic acids research (2014): gku1158.
10. Andrade, Sónia CS, et al. "A transcriptomic approach to ribbon worm systematics (Nemertea): resolving the Pilidiophora problem." Molecular biology and evolution 31.12 (2014): 3206-3215.
11. Berglund, Ann-Charlotte, et al. "InParanoid 6: eukaryotic ortholog clusters with inparalogs." Nucleic acids research 36.suppl 1 (2008): D263-D266.
12. Børve, Aina, and Andreas Hejnol. "Development and juvenile anatomy of the nemertodermatid Meara stichopi (Bock) Westblad 1949 (Acoelomorpha)." Frontiers in zoology 11.1 (2014): 1.

13. Bourlat, S J. et al. (2006) Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida Nature 12/2006; 444(7115):85-8. DOI:10.1038/nature05241
14. Bourlat, Sarah J., et al. "The mitochondrial genome structure of Xenoturbella bocki (phylum Xenoturbellida) is ancestral within the deuterostomes." BMC Evolutionary Biology 9.1 (2009): 1.
15. Bourlat, Sarah J., et al. "Xenoturbella is a deuterostome that eats molluscs." Nature 424.6951 (2003): 925-928.
16. Bradnam, Keith R., et al. "Assemblathon 2: evaluating de novo methods of genome assembly in three vertebrate species." GigaScience 2.1 (2013): 1-31.
17. Brady, Arthur, and Steven Salzberg. "PhymmBL expanded: confidence scores, custom databases, parallelization and more." Nature methods 8.5 (2011): 367-367.
18. Brejová, Broňa, et al. "Finding genes in Schistosoma japonicum: annotating novel genomes with help of extrinsic evidence." Nucleic acids research (2009): gkp052.
19. Burge, Chris, and Samuel Karlin. "Prediction of complete gene structures in human genomic DNA." Journal of molecular biology 268.1 (1997): 78-94.
20. Cannon, Johanna Taylor, et al. "Xenacoelomorpha is the sister group to Nephrozoa." Nature 530.7588 (2016): 89-93.
21. Capella-Gutiérrez, Salvador, José M. Silla-Martínez, and Toni Gabaldón. "trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses." Bioinformatics 25.15 (2009): 1972-1973.
22. Chinwalla, Asif T., et al. "Initial sequencing and comparative analysis of the mouse genome." Nature 420.6915 (2002): 520-562.
23. Compeau, Phillip EC, Pavel A. Pevzner, and Glenn Tesler. "How to apply de Bruijn graphs to genome assembly." Nature biotechnology 29.11 (2011): 987-991.
24. Cook, Charles E., et al. "The Hox gene complement of acoel flatworms, a basal bilaterian clade." Evolution & development 6.3 (2004): 154-163.
25. Davis, Matthew PA, et al. "Kraken: a set of tools for quality control and analysis of high-throughput sequence data." Methods 63.1 (2013): 41-49.
26. De Robertis, E. M. "The molecular ancestry of segmentation mechanisms." Proceedings of the National Academy of Sciences 105.43 (2008): 16411-16412.
27. Dehal, Paramvir, et al. "The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins." Science 298.5601 (2002): 2157-2167.
28. Dessimoz, C. et al. (2005) OMA, a comprehensive, automated project for the identification of orthologs from complete genome data: Introduction and first

achievements. In Aoife McLysath and Daniel H. Huson, editors, RECOMB 2005 Workshop on Comparative Genomics, volume LNBI 3678 of Lecture Notes in Bioinformatics, pages 61–72. Springer-Verlag, 2005.

29. Dessimoz, Christophe, et al. "Detecting non-orthology in the COGs database and other approaches grouping orthologs using genome-specific best hits."Nucleic acids research 34.11 (2006): 3309-3316.
30. Dessimoz, Christophe, et al. "Fast estimation of the difference between two PAM/JTT evolutionary distances in triplets of homologous sequences." BMC bioinformatics 7.1 (2006): 529.
31. Dörjes, Jürgen. "Die Acoela (Turbellaria) der Deutschen Nordseeküste." Journal of Zoological Systematics and Evolutionary Research 6.2-4 (1968): 57-452.
32. Doyle, Jeff J. "Gene trees and species trees: molecular systematics as one-character taxonomy." Systematic Botany (1992): 144-163.
33. Dunn, Casey W., et al. "Broad phylogenomic sampling improves resolution of the animal tree of life." Nature 452.7188 (2008): 745-749.
34. Edgar, Robert C. "MUSCLE: multiple sequence alignment with high accuracy and high throughput." Nucleic acids research 32.5 (2004): 1792-1797.
35. Edgar, Robert C. "Search and clustering orders of magnitude faster than BLAST." Bioinformatics 26.19 (2010): 2460-2461.
36. Egger, Bernhard, et al. "A Transcriptomic-Phylogenomic Analysis of the Evolutionary Relationships of Flatworms." Current Biology 25.10 (2015): 1347-1353.
37. Egger, Bernhard, Robert Gschwentner, and Reinhard Rieger. "Free-living flatworms under the knife: past and present." Development genes and evolution 217.2 (2007): 89-104.
38. Ehlers, Ulrich. "Frontal glandular and sensory structures inNemertoderma (Nemertodermatida) and Paratomella (Acoela): ultrastructure and phylogenetic implications for the monophyly of the Euplathelminthes (Plathelminthes)." Zoomorphology 112.4 (1992): 227-236.
39. Ehlers, Ulrich. “Phylogenetische System der Plathelminthes.” G. Fischer, 1985.
40. Fernández, Rosa, Gustavo Hormiga, and Gonzalo Giribet. "Phylogenomic analysis of spiders reveals nonmonophyly of orb weavers." Current Biology24.15 (2014): 1772-1777.
41. Ferrier, David EK. "Evolution of Hox gene clusters." Hox gene expression. Springer New York, 2007. 53-67.
42. Flicek, Paul, et al. "Ensembl 2012." Nucleic acids research (2011): gkr991.

43. Franzén, Å., Afzelius, B.A. 1987. The ciliated epidermis of Xenoturbella bocki (Platyhelminthes, Xenoturbellida) with some phylogenetic considerations. Zool. Scripta 16: 9-17.
44. Friedman, Robert, and Austin L. Hughes. "Pattern and timing of gene duplication in animal genomes." Genome research 11.11 (2001): 1842-1847.
45. Fritzsch G, Bohme MU, Thorndyke M, Nakano H, Israelsson O, Stach T, Schlegel M, Hankeln T, Stadler PF. (2007). A PCR Survey of Xenoturbella bocki Hox Genes. J Exp Zool B Mol Dev Evol 310B:278–284.
46. Fu, Limin, et al. "CD-HIT: accelerated for clustering the next-generation sequencing data." Bioinformatics 28.23 (2012): 3150-3152.
47. Fulton, D L. et al. (2006) Improving the specificity of high-throughput ortholog prediction. BMC Bioinformatics, 28(7): 270
48. Gee H. 2003. "You aren't what you eat." Nature 424: 885–886.
49. Gerhart, John, Christopher Lowe, and Marc Kirschner. "Hemichordates and the origin of chordates." Current opinion in genetics & development 15.4 (2005): 461-467.
50. Gonnet, Gaston H., Mark A. Cohen, and Steven A. Benner. "Exhaustive matching of the entire protein sequence database." Science 256.5062 (1992): 1443-1445.
51. Grabherr, Manfred G., et al. "Full-length transcriptome assembly from RNA-Seq data without a reference genome." Nature biotechnology 29.7 (2011): 644-652.
52. Graff, L. von. "Turbellaria, Acoela und Rhabdocoelida." (1904).
53. Gregory, T. Ryan. "Genome size evolution in animals." The evolution of the genome 1 (2005): 4-87.
54. Guigó, Roderic, et al. "EGASP: the human ENCODE genome annotation assessment project." Genome Biol 7.Suppl 1 (2006): S2.
55. Guindon, Stéphane, et al. "New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0." Systematic biology 59.3 (2010): 307-321.
56. Haas, Alexander. "Phylogeny of frogs as inferred from primarily larval characters (Amphibia: Anura)." Cladistics 19.1 (2003): 23-89.
57. Haeckel, Ernst. "Memoirs: the Gastraea-Theory, the phylogenetic classification of the animal kingdom and the homology of the germ-lamellæ." Journal of Cell Science 2.54 (1874): 142-165.
58. Haszprunar, G. "Plathelminthes and Plathelminthomorpha—paraphyletic taxa." Journal of Zoological Systematics and Evolutionary Research 34.1 (1996): 41-48.

59. Hejnol, Andreas, and Mark Q. Martindale. "Acoel development supports a simple planula-like urbilaterian." Philosophical Transactions of the Royal Society of London B: Biological Sciences 363.1496 (2008): 1493-1501.
60. Hejnol, Andreas, et al. "Assessing the root of bilaterian animals with scalable phylogenomic methods." Proceedings of the Royal Society of London B: Biological Sciences 276.1677 (2009): 4261-4270.
61. Hendelberg, J. "Comparative morphology of turbellarian spermatozoa studied by electron microscopy." Acta zool. fenn 154 (1977): 149-162.
62. Henikoff,J.G., Peitrokovski,S. and Henikoff,S. (1997) Recent enhancements to the Blocks database servers. Nucleic Acids Res., 25, 222–225.
63. Hoff, Katharina J., and Mario Stanke. "WebAUGUSTUS—a web service for training AUGUSTUS and predicting genes in eukaryotes." Nucleic acids research 41.W1 (2013): W123-W128.
64. Hooge, Matthew D., et al. "Molecular systematics of the Acoela (Acoelomorpha, Platyhelminthes) and its concordance with morphology." Molecular phylogenetics and evolution 24.2 (2002): 333-342.
65. Hou, Yubo, and Senjie Lin. "Distinct gene number-genome size relationships for eukaryotes and non-eukaryotes: gene content estimation for dinoflagellate genomes." PLoS One 4.9 (2009): e6978.
66. Huerta-Cepas et al. (2014). "PhylomeDB v4: zooming into the plurality of evolutionary histories of a genome.". Nucleic Acids Res. (England) 42 (Database issue): D897–902. doi:10.1093/nar/gkt1177
67. Huerta-Cepas, Jaime, et al. "PhylomeDB: a database for genome-wide collections of gene phylogenies." Nucleic acids research 36.suppl 1 (2008): D491-D496.
68. Hughes, Austin L., and Robert Friedman. "Differential loss of ancestral gene families as a source of genomic divergence in animals." Proceedings of the Royal Society of London B: Biological Sciences 271.Suppl 3 (2004): S107-S109.
69. Hyman, Libbie Henrietta. "The invertebrates: smaller coelomate groups, Chaetognatha, Hemi-chordata, Pogonophora, Phoronida, Ectoprocta, Brachipoda, Sipunculida, the coelomate Bila-teria. Volume V." The invertebrates: smaller coelomate groups, Chaetognatha, Hemi-chordata, Pogonophora, Phoronida, Ectoprocta, Brachipoda, Sipunculida, the coelomate Bila-teria. Volume V. (1959).
70. Israelsson, Olle. "New light on the enigmatic Xenoturbella (phylum uncertain): ontogeny and phylogeny." Proceedings of the Royal Society of London B: Biological Sciences 266.1421 (1999): 835-841.

71. Jägersten, Gösta. On the early phylogeny of the Metazoa: the bilaterogastraea theory. Almqvist & Wiksells Boktr., 1955.
72. Jennings, J. B. 1971. Parasitism and commensalism in the Turbellaria. Adv. ParasitoL 9: 1-32.
73. Jondelius, U. et al. 2002: The Nemertodermatida are basal bilaterians and not members of the Platyhelminthes. Zoologica scripta, 31: 201-215.
74. Jondelius, Ulf, et al. "How the worm got its pharynx: phylogeny, classification and Bayesian assessment of character evolution in Acoela." Systematic Biology (2011): syr073.
75. Karger, David R., and Clifford Stein. "A new approach to the minimum cut problem." Journal of the ACM (JACM) 43.4 (1996): 601-640.
76. Karger, David R., Philip N. Klein, and Robert E. Tarjan. "A randomized linear-time algorithm to find minimum spanning trees." Journal of the ACM (JACM)42.2 (1995): 321-328.
77. Kersey, Paul J., et al. "Ensembl Genomes: extending Ensembl across the taxonomic space." Nucleic acids research 38.suppl 1 (2010): D563-D569.
78. Kersey, Paul, et al. "Integr8 and Genome Reviews: integrated views of complete genomes and proteomes." Nucleic acids research 33.suppl 1 (2005): D297-D302.
79. Kimmel, CB "Was Urbilateria segmented?" Trends Genet. 1996 Sep;12(9):329-31.
80. Kocot, Kevin M., et al. "Phylogenomics reveals deep molluscan relationships."Nature 477.7365 (2011): 452-456.
81. Kriventseva, Evgenia V., et al. "OrthoDB: the hierarchical catalog of eukaryotic orthologs." Nucleic acids research 36.suppl 1 (2008): D271-D275.
82. Kumar, Sujai, et al. "Blobology: exploring raw genome data for contaminants, symbionts and parasites using taxon-annotated GC-coverage plots."Frontiers in genetics 4 (2013): 237.
83. Langmead, Ben, and Steven L. Salzberg. "Fast gapped-read alignment with Bowtie 2." Nature methods 9.4 (2012): 357-359.
84. Lartillot, Nicolas, and Hervé Philippe. "A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process." Molecular biology and evolution 21.6 (2004): 1095-1109.
85. Lartillot, Nicolas, and Hervé Philippe. "Improvement of molecular phylogenetic inference and the phylogeny of Bilateria." Philosophical Transactions of the Royal Society B: Biological Sciences 363.1496 (2008): 1463-1472.

86. Lartillot, Nicolas, et al. "PhyloBayes MPI. Phylogenetic reconstruction with infinite mixtures of profiles in a parallel environment." Systematic biology(2013): syt022.
87. Lartillot, Nicolas, Henner Brinkmann, and Hervé Philippe. "Suppression of long-branch attraction artefacts in the animal phylogeny using a site-heterogeneous model." BMC evolutionary biology 7.1 (2007): 1.
88. Lartillot, Nicolas, Thomas Lepage, and Samuel Blanquart. "PhyloBayes 3: a Bayesian software package for phylogenetic reconstruction and molecular dating." Bioinformatics 25.17 (2009): 2286-2288.
89. Laumer, Christopher E., Andreas Hejnol, and Gonzalo Giribet. "Nuclear genomic signals of the 'microturbellarian'roots of platyhelminth evolutionary innovation." Elife 4 (2015): e05503.
90. Laumer, Christopher E., et al. "Spiralian phylogeny informs the evolution of microscopic lineages." Current Biology 25.15 (2015): 2000-2006.
91. Lemmon, Alan R., et al. "The effect of ambiguous data on phylogenetic estimates obtained by maximum likelihood and Bayesian inference." Systematic Biology 58.1 (2009): 130-145.
92. Li Y, et al. (2014) Genomic Evolution of Saccharomyces cerevisiae under Chinese Rice Wine Fermentation. Genome Biol Evol 6(9):2516-26
93. Li, Li, Christian J. Stoeckert, and David S. Roos. "OrthoMCL: identification of ortholog groups for eukaryotic genomes." Genome research 13.9 (2003): 2178-2189.
94. Littlewood, D. T. J., P. D. Olson, M. J. Telford, E. A. Herniou, and M. Riutort. 2001. Elongation factor 1-alpha sequences alone do not assist in resolving the position of the acoela within the metazoa. Mol. Biol. Evol. 18:437–442.
95. Lundin K. 2001. Degenerating epidermal cells in Xenoturbella bocki (phylum uncertain). Nemertodermatida and Acoela (Platyhelminthes). Belgian J Zool 131:153–157.
96. Lundin, K. (1998). The epidermal ciliary rootlets of Xenoturbella bocki (Xenoturbellida) revisited: new support for a possible kinship with the Acoelomorpha (Platyhelminthes). Zoologica Scripta, 27, 263–270.
97. Luo, Ruibang, et al. "SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler." Gigascience 1.1 (2012): 18.
98. Madden, Thomas. "The BLAST sequence analysis tool." (2013).
99. Marçais, Guillaume, and Carl Kingsford. "A fast, lock-free approach for efficient parallel counting of occurrences of k-mers." Bioinformatics 27.6 (2011): 764-770.

100. Marlétaz, Ferdinand, et al. "Chaetognath phylogenomics: a protostome with deuterostome-like development." Current Biology 16.15 (2006): R577-R578.
101. Martín-Durán, José M., Francisco Monjo, and Rafael Romero. "Planarian embryology in the era of comparative developmental biology." International Journal of Developmental Biology 56.1-2-3 (2012): 39-48.
102. Martín-Durán, José María, and Bernhard Egger. "Developmental diversity in free-living flatworms." EvoDevo 3.1 (2012): 1-23.
103. Matus, David Q., et al. "Molecular evidence for deep evolutionary roots of bilaterality in animal development." Proceedings of the National Academy of Sciences 103.30 (2006): 11195-11200.
104. McLysaght, Aoife, Karsten Hokamp, and Kenneth H. Wolfe. "Extensive genomic duplication during early chordate evolution." Nature genetics 31.2 (2002): 200-204.
105. McLysaght, Aoife, Karsten Hokamp, and Kenneth H. Wolfe. "Extensive genomic duplication during early chordate evolution." Nature genetics 31.2 (2002): 200-204.
106. Medvedev, Paul, et al. "Computability of models for sequence assembly."International Workshop on Algorithms in Bioinformatics. Springer Berlin Heidelberg, 2007.
107. Merchant, Samier, Derrick E. Wood, and Steven L. Salzberg. "Unexpected cross-species contamination in genome sequencing projects." PeerJ 2 (2014): e675.
108. Mewes, HrW, et al. "Overview of the yeast genome." Nature 387.6632 (1997): 7-8.
109. Mi, Huaiyu, et al. "PANTHER version 6: protein sequence and function evolution data with expanded representation of biological pathways." Nucleic acids research 35.suppl 1 (2007): D247-D252.
110. Mitchelson, Keith R., ed. New high throughput technologies for DNA sequencing and genomics. Vol. 2. Elsevier, 2011.
111. Muller, Jean, et al. "eggNOG v2. 0: extending the evolutionary genealogy of genes with enhanced non-supervised orthologous groups, species and functional annotations." Nucleic acids research 38.suppl 1 (2010): D190-D195.
112. Nakano, H. et al. (2013). "Xenoturbella bocki exhibits direct development with similarities to Acoelomorpha". Nature Communications 4: 1537.doi:10.1038/ncomms2556. PMC 3586728.PMID 23443565.
113. Nielsen, Rodney D., et al. "A taxonomy of questions for question generation."Proceedings of the Workshop on the Question Generation Shared Task and Evaluation Challenge. 2008.
114. Nordborg, M. & Tavaré, S. Linkage disequilibrium: what history has to tell us. Trends Genet. 18, 83–90 (2003)

115. Nordborg, Magnus, et al. "The pattern of polymorphism in Arabidopsis thaliana."PLoS biology 3.7 (2005): 1289.
116. Norén, Michael, and Ulf Jondelius. "Xenoturbella's molluscan relativest." Nature390 (1997): 31-32.
117. Ohta, M., K. Ina, K. Kusuzaki, N. Kido, Y. Arakawa, N. Kato 1991. Cloning and expression of the rfe-rff gene cluster of Escherichia coli. Mol.Microbiol. 5:1853-1862
118. Östlund, Gabriel, et al. "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." Nucleic acids research 38.suppl 1 (2010): D196-D203.
119. Östlund, Gabriel, et al. "InParanoid 7: new algorithms and tools for eukaryotic orthology analysis." Nucleic acids research 38.suppl 1 (2010): D196-D203.
120. Paps, Jordi, Jaume Baguñà, and Marta Riutort. "Bilaterian phylogeny: a broad sampling of 13 nuclear genes provides a new Lophotrochozoa phylogeny and supports a paraphyletic basal Acoelomorpha." Molecular Biology and Evolution26.10 (2009): 2397-2406.
121. Parra, Genis, Keith Bradnam, and Ian Korf. "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes." Bioinformatics 23.9 (2007): 1061-1067.
122. Parra, Genis, Keith Bradnam, and Ian Korf. "CEGMA: a pipeline to accurately annotate core genes in eukaryotic genomes." Bioinformatics 23.9 (2007): 1061-1067.
123. Perea-Atienza, Elena, et al. "The nervous system of Xenacoelomorpha: a genomic perspective." Journal of Experimental Biology 218.4 (2015): 618-628.
124. Perseke, Marleen, et al. "Evolution of mitochondrial gene orders in echinoderms." Molecular phylogenetics and evolution 47.2 (2008): 855-864.
125. Perseke, Marleen, et al. "Evolution of mitochondrial gene orders in echinoderms." Molecular phylogenetics and evolution 47.2 (2008): 855-864.
126. Peterson, Kevin J., and Douglas J. Eernisse. "Animal phylogeny and the ancestry of bilaterians: inferences from morphology and 18S rDNA gene sequences." Evolution & development 3.3 (2001): 170-205.
127. Philippe H, Brinkmann H, Lavrov DV, Littlewood DTJ, Manuel M, Wörheide G, et al. (2011) Resolving Difficult Phylogenetic Questions: Why More Sequences Are Not Enough. PLoS Biol 9(3): e1000602. doi:10.1371/journal.pbio.1000602
128. Philippe H, Brinkmann H, Martinez P, Riutort M, Baguñà J (2007) Acoel Flatworms Are Not Platyhelminthes: Evidence from Phylogenomics. PLoS ONE 2(8): e717. doi:10.1371/journal.pone.0000717
129. Philippe, Herve, and Maximilian J. Telford. "Large-scale sequencing and the new animal phylogeny." Trends in Ecology & Evolution 21.11 (2006): 614-620.

130. Philippe, Hervé, et al. "Acoelomorph flatworms are deuterostomes related to Xenoturbella." Nature 470.7333 (2011): 255-258.
131. Philippe, Hervé, et al. "Acoelomorph flatworms are deuterostomes related to Xenoturbella." Nature 470.7333 (2011): 255-258.
132. Philippe, Hervé, et al. "Phylogenomics revives traditional views on deep animal relationships." Current Biology 19.8 (2009): 706-712.
133. Philippe, Hervé, et al. "Phylogenomics revives traditional views on deep animal relationships." Current Biology 19.8 (2009): 706-712.
134. Philippe, Hervé, et al. "Phylogenomics." Annual Review of Ecology, Evolution, and Systematics (2005): 541-562.
135. Pick, K. S., et al. "Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships." Molecular biology and evolution 27.9 (2010): 1983-1987.
136. Pick, K. S., et al. "Improved phylogenomic taxon sampling noticeably affects nonbilaterian relationships." Molecular biology and evolution 27.9 (2010): 1983-1987.
137. Podsiadlowski, Lars, et al. "Phylogeny and mitochondrial gene order variation in Lophotrochozoa in the light of new mitogenomic data from Nemertea." BMC genomics 10.1 (2009): 364.
138. Price, Alkes L., et al. "Principal components analysis corrects for stratification in genome-wide association studies." Nature genetics 38.8 (2006): 904-909.
139. Pruitt, Kim D., Tatiana Tatusova, and Donna R. Maglott. "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." Nucleic acids research 35.suppl 1 (2007): D61-D65.
140. Pryszcz, L P. et al. (2011) MetaPhOrs: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Res. 39: e32.
141. Putnam, N. H. et al. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. Science 317, 86–94 (2007)
142. Raikova, O. I. et al. (2000). An immunocytochemical and ultrastructural study of the nervous and muscular systems of Xenoturbella westbladi (Bilateria inc. sed.). Zoomorphology, 120, 107–118.
143. Ramirez-Gonzalez R. (2013). Kontaminant, a k-mer based contamination screening and filtering tool. Available online at: http://www.tgac.ac.uk/kontaminant
144. Remm, M. et al (2001) Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol, 314(5):1041–52

145. Rieger, R. M., S. Tyler, J. P. S. Smith, and G. Rieger. 1991. Platyhelminthes: Turbellaria. In F. W. Harrison and B. J. Bogitsh (eds.), Microscopic anatomy of invertebrates, pp. 7–140. Wiley-Liss, New York

146. Rieger, Reinhard M., and Peter Ladurner. "The significance of muscle cells for the origin of mesoderm in Bilateria." Integrative and comparative biology 43.1 (2003): 47-54.

147. Rohde K, Watson N, Cannon LRG. 1988. Ultrastructure of epidermal cilia of Pseudactinoposthia sp. (Platyhelminthes Acoela) - implications for the phylogenetic status of the Xenoturbellida and Acoelomorpha. J Submicrosc Cytol Pathol 20:759–767.

148. Roth, Alexander CJ, Gaston H. Gonnet, and Christophe Dessimoz. "Algorithm of OMA for large-scale orthology inference." BMC bioinformatics 9.1 (2008): 518.

149. Roure B, Baurain D, Philippe H. Impact of missing data on phylogenies inferred from empirical phylogenomic datasets. Mol Biol Evol. 2013;30:197–214.

150. Roure, Béatrice, Denis Baurain, and Hervé Philippe. "Impact of missing data on phylogenies inferred from empirical phylogenomic data sets." Molecular biology and evolution 30.1 (2013): 197-214.

151. Rouse, Greg W., et al. "New deep-sea species of Xenoturbella and the position of Xenacoelomorpha." Nature 530.7588 (2016): 94-97.

152. Ruiz-Trillo, I., M. Riutort, D. T. J. Litlewood, E. A. Herniou, and J. Baguñà. 1999. Acoel flatworms: Earliest extant bilaterian metazoans, not members of Platyhelminthes. Science, 2831919-1923.

153. Ruiz-Trillo, Iñaki, et al. "A phylogenetic analysis of myosin heavy chain type II sequences corroborates that Acoela and Nemertodermatida are basal bilaterians." Proceedings of the National Academy of Sciences 99.17 (2002): 11246-11251.

154. Ruiz-Trillo, Iñaki, et al. "Mitochondrial genome data support the basal position of Acoelomorpha and the polyphyly of the Platyhelminthes." Molecular phylogenetics and evolution 33.2 (2004): 321-332.

155. Salichos, Leonidas, and Antonis Rokas. "Evaluating ortholog prediction algorithms in a yeast model clade." (2011): e18755.

156. Salzberg, Steven L., and James A. Yorke. "Bioinformatics Letter To The Editor." Bioinformatics 21.24 (2005): 4320-4321.

157. Schmieder, Robert, and Robert Edwards. "Quality control and preprocessing of metagenomic datasets." Bioinformatics 27.6 (2011): 863-864.

158. Schneider, Adrian, Christophe Dessimoz, and Gaston H. Gonnet. "OMA Browser—exploring orthologous relations across 352 complete genomes."Bioinformatics 23.16 (2007): 2180-2182.

159. Semmler, Henrike, Xavier Bailly, and Andreas Wanninger. "Myogenesis in the basal bilaterian Symsagittifera roscoffensis (Acoela)." Frontiers in zoology 5.1 (2008): 1.

160. Sempere, L. F., Cole, C. N., McPeek, M. A., and Peterson, K. J. 2006. The phylogenetic distribution of metazoan microRNAs: insights into evolutionary complexity and constraint. J. Exp. Zool. (Mol. Dev. Evol.) 306B: 575–588.

161. Sempere, L. F., Martinez, P., Cole, C., Bagun˜a`, J., and Peterson, K. J. 2007. Phylogenetic distribution of microRNAs supports the basal position of acoel flatworms and the polyphyly of Platyhelminthes. Evol. Dev. 9: 409–415

162. Shrestha, Anish Man Singh, Martin C. Frith, and Paul Horton. "A bioinformatician's guide to the forefront of suffix array construction algorithms."Briefings in bioinformatics 15.2 (2014): 138-154.

163. Simakov, Oleg, et al. "Insights into bilaterian evolution from three spiralian genomes." Nature 493.7433 (2013): 526-531.

164. Simakov, Oleg, et al. "Insights into bilaterian evolution from three spiralian genomes." Nature 493.7433 (2013): 526-531.

165. Simpson, Jared T. "Exploring genome characteristics and sequence quality without a reference." Bioinformatics (2014): btu023.

166. Simpson, Jared T. "Genome informatics 2014." Genome biology 15.11 (2014): 1.

167. Simpson, Jared T., et al. "ABySS: a parallel assembler for short read sequence data." Genome research 19.6 (2009): 1117-1123.

168. Smith, J. P. S., Tyler, S. (1986). Frontal organs in the Acoelomorpha (Turbellaria): ultrastructure and phylogenetic significance. Hydrobiologia, 132, 71–78.

169. Smith, Stephen A., et al. "Resolving the evolutionary relationships of molluscs with phylogenomic tools." Nature 480.7377 (2011): 364-367.

170. Smith, Temple F., and Michael S. Waterman. "Identification of common molecular subsequences." Journal of molecular biology 147.1 (1981): 195-197.

171. Sodergren, Erica, et al. "The genome of the sea urchin Strongylocentrotus purpuratus." Science 314.5801 (2006): 941-952.

172. Srivastava, Mansi, et al. "Whole-body acoel regeneration is controlled by Wnt and Bmp-Admp signaling." Current Biology 24.10 (2014): 1107-1113.

173. Stach, E. M., and Alan T. Bull. "Estimating and comparing the diversity of marine actinobacteria." Antonie van Leeuwenhoek 87.1 (2005): 3-9.

174. Stamatakis, Alexandros. "RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies." Bioinformatics 30.9 (2014): 1312-1313.
175. Stanke, Mario, et al. "Gene prediction in eukaryotes with a generalized hidden Markov model that uses hints from external sources." BMC bioinformatics 7.1 (2006): 62.
176. Steinbock, O. T. T. O. "Marine Turbellaria." Zoology of the Faroes VIII (1930): 1-26.
177. Steinbock, O. T. T. O. "The Zoology of Iceland." Vol. II, Part 9. Copenh. a. Reyk. (1938)
178. Struck, Torsten H. "Direction of evolution within Annelida and the definition of Pleistoannelida." Journal of Zoological Systematics and Evolutionary Research49.4 (2011): 340-345.
179. Struck, Torsten H., and Frauke Fisse. "Phylogenetic position of Nemertea derived from phylogenomic data." Molecular biology and evolution 25.4 (2008): 728-736.
180. Tarailo Graovac, Maja, and Nansheng Chen. "Using RepeatMasker to identify repetitive elements in genomic sequences." Current Protocols in Bioinformatics(2009): 4-10.
181. Tatusov, Roman L., et al. "The COG database: an updated version includes eukaryotes." BMC bioinformatics 4.1 (2003): 41.
182. Tekle, Yonas I., et al. "Revision of the Childiidae (Acoela), a total evidence approach in reconstructing the phylogeny of acoels with reversed muscle layers." Journal of Zoological Systematics and Evolutionary Research 43.1 (2005): 72-90.
183. Telford, M J. 2008. Xenoturbellida: The fourth deuterostome phylum and the diet of worms. genesis 11/2008; 46(11):580-6. DOI:10.1002/dvg.20414
184. Telford, M. J. 2000. Evidence for the derivation of the Drosophila fushi tarazu gene from a Hox gene orthologous to lophotrochozoan Lox5. Curr. Biol, 10349-352.
185. Telford, Maximilian J. "The Animal Tree of Life." Science 339.6121 (2013): 764-766.
186. Telford, Richard D., et al. "Footstrike is the major cause of hemolysis during running." Journal of Applied Physiology 94.1 (2003): 38-42.
187. Tempel, Sébastien. "Using and understanding RepeatMasker." Mobile Genetic Elements: Protocols and Genomic Applications (2012): 29-51.
188. Todt, Christiane. "Structure and evolution of the pharynx simplex in acoel flatworms (Acoela)." Journal of morphology 270.3 (2009): 271-290.
189. Trivedi, Urmi H., et al. "Quality control of next-generation sequencing data without a reference." Frontiers in genetics 5 (2014).
190. Tyler, S., & Rieger, R. M. (1977). Ultrastructural evidence for the systematic position of the Nemertodermatida (Turbellaria). Acta Zoologica Fennica, 194–207.

191. UniProt Consortium. "The universal protein resource (UniProt) in 2010." Nucleic acids research 38.suppl 1 (2010): D142-D148.
192. Van Dongen, Stijn Marinus. "Graph clustering by flow simulation." (2000).
193. Venter, J. Craig, et al. "The sequence of the human genome." science291.5507 (2001): 1304-1351.
194. Vinson, Jade P., et al. "Assembly of polymorphic genomes: algorithms and application to Ciona savignyi." Genome research 15.8 (2005): 1127-1135.
195. von Reumont, Bjoern M., et al. "Pancrustacean phylogeny in the light of new phylogenomic data: support for Remipedia as the possible sister group of Hexapoda." Molecular Biology and Evolution 29.3 (2012): 1031-1045.
196. Wallberg, A. et al. (2007). "Dismissal of Acoelomorpha: Acoela and Nemertodermatida are separate early bilaterian clades". Zoologica Scripta 36 (5): 509–523. doi:10.1111/j.1463-6409.2007.00295.x.
197. Westblad, E. "STUDIEN BER SKANDINAVISCHE TURBELLARIA ACOELA."Studien über skandinavische Turbellaria Acoela (1937): 191-273.
198. Westblad, E. (1949) Xenoturbella bocki n. g., n. sp., a peculiar, primitive Turbellarian type. Arkiv för Zoologi 1:3-29
199. Westheide, W., & Rieger, R. M. (2007). Systematik-Poster: Zoologie. Heidelberg: Spektrum Akademischer Verlag.
200. Wood, Derrick E., and Steven L. Salzberg. "Kraken: ultrafast metagenomic sequence classification using exact alignments." Genome Biol 15.3 (2014): R46.
201. Xavier-Neto, J., et al. "Parallel avenues in the evolution of hearts and pumping organs." Cell Mol Life Sci 64.6 (2007): 719-734.
202. Xie, Yinlong, et al. "SOAPdenovo-Trans: de novo transcriptome assembly with short RNA-Seq reads." Bioinformatics 30.12 (2014): 1660-1666.
203. Xu, Yizhuang, et al. "PASA–a program for automated protein NMR backbone signal assignment by pattern-filtering approach." Journal of biomolecular NMR34.1 (2006): 41-56.
204. Zerbino DR, Birney E. Velvet: algorithms for de novo short read assembly using de Bruijn graphs.Genome Res. 2008;18:821–829.
205. Zhang, Guofan, et al. "The oyster genome reveals stress adaptation and complexity of shell formation." Nature 490.7418 (2012): 49-54.
206. Zrzavý, Jan, et al. "Phylogeny of the Metazoa based on morphological and 18S ribosomal DNA evidence." Cladistics 14.3 (1998): 249-285.

List of definitions and abbreviations

ABySS – Assembly By Short Sequences; de novo short read assembly algorithm (Simpson et al. 2009).

AIC - Akaike Information Criterion.

Assemblathon 2 - The Assemblathon competitions are intended to assess current state-of-the-art methods in genome assembly (Bradnam at al. 2013).

AUGUSTUS - Predicts genes in eukaryotic genomic sequences. AUGUSTUS is based on the evaluation of hints to potentially protein-coding regions by means of a Generalized Hidden Markov Model (GHMM) that takes both intrinsic and extrinsic information into account (Stanke et al. 2008).

Baysian tree reconstruction method - Bayesian inference of phylogeny uses a likelihood function to create a quantity called the posterior probability of trees using a model of evolution, based on some prior probabilities, producing the most likely phylogenetic tree for the given data.

BLAST - BLAST for Basic Local Alignment Search Tool

BLOSUM - BLOcks SUbstitution Matrix

bowtie2 – An ultrafast and memory-efficient tool for aligning sequencing reads to long reference sequences.

CAT+GTR+ Γ model - A Bayesian Mixture Model for Across-Site Heterogeneities in the Amino-Acid Replacement Process.

CEGMA – Core Eukaryotic Genes Mapping Approach, the algorithm which uses Hidden Markov probabilistic gene models for finding orthologous genes in the proteomes

EggNOG - A database of orthologous groups and functional annotation (http://eggnogdb.embl.de/#/app/home).

Ensembl Compara - Ensembl Compara provides cross-species resources and analyses, at both the sequence level and the gene level. These resources are described in more details in Herrero et al., Database, 2016.

ESPRIT software - Comparative genomics approach to detecting split-coding regions in a low-coverage genome.

ESTs – Expressed Sequence Tags

FASTA - FASTA format is a text-based format for representing either nucleotide sequences or peptide sequences.

FastQC - Modern high throughput sequencers can generate tens of millions of sequences in a single run. Before analysing this sequence to draw biological conclusions you should always perform some simple quality control checks to ensure that the raw data looks good and there are no problems or biases in your data which may affect how you can usefully use it.

Gbp – Giga base pair.

GenScan - Identifies complete exon/intron structures of genes in genomic DNA. Features of the program include the capacity to predict multiple genes in a sequence, to deal with partial as well as complete genes, and to predict consistent sets of genes occurring on either or both DNA strands. Genscan is one of the best gene finding algorithms. The UCSC Genome Browser (http://genome.ucsc.edu) is a convenient graphic visualization tool for genome annotations.

GTR+ Γ model - Reversible model of nucleotide substitution under the Gamma model of rate heterogeneity.

HMM (Hidden Markov Model) - A hidden Markov model (HMM) is a statistical Markov model in which the system being modeled is assumed to be a Markov process with unobserved (hidden) states.

HMMER - HMMER is a free and commonly used software package for sequence analysis written by Sean Eddy. Its general usage is to identify homologous protein or nucleotide sequences.

HOGs - Hierarchical orthologous groups are sets of genes that are defined with respect to particular taxonomic ranges of interest. They group genes that have descended from a single common ancestral genes in that taxonomic range.

htgs database – The High Throughput Genomic database, contains contigs greater than 2 kb from genomic sequence projects which are made available to the scientific community before their publication

Illumina technology - https://www.illumina.com/content/dam/illumina-marketing/documents/products/illumina_sequencing_introduction.pdf

InParanoid

Jellyfish multithreaded k-mer counter

JTT substitution matrix - The Dayhoff PAM matrices were based on relatively few alignments (since not more were available at that time), but in the 1990s, new matrices were estimated using almost the same methodology, but based on the large protein databases available then (the latter being known as "JTT" matrices).

Kbp – Kilo base pair

LG substitution matrix - An Improved General Amino Acid Replacement Matrix (Le & Gascuel 2008)

Mann-Whitney U test - Mann–Whitney U test (also called the Mann–Whitney–Wilcoxon (MWW), Wilcoxon rank-sum test, or Wilcoxon–Mann–Whitney test) is a nonparametric test of the null hypothesis that it is equally likely that a randomly selected value from one sample will be less than or greater than a randomly selected value from a second sample.

Maximum likelihood method - maximum likelihood estimation (MLE) is a method of estimating the parameters of a statistical model given observations, by finding the parameter values that maximize the likelihood of making the observations given the parameters.

Maximum parsimony method - Maximum parsimony predicts the evolutionary tree or trees that minimize the number of steps required to generate the observed variation in the sequences from common ancestral sequences. For this reason, the method is also sometimes referred to as the minimum evolution method.

Mbp – Mega base pair

MCL - Markov Clustering

MCMC - Markov chain Monte Carlo (MCMC) methods are a class of algorithms for sampling from a probability distribution based on constructing a Markov chain that has the desired distribution as its equilibrium distribution.

ML – Maximum Likelihood

MUSCLE - MUltiple Sequence Comparison by Log- Expectation. MUSCLE is claimed to achieve both better average accuracy and better speed than ClustalW2 or T-Coffee, depending on the chosen options.

NCBI – National Center for Biotechnology Information

NGS – Next Generation Sequencing

nr database – non-redundant sequence database, contains non-redundant sequences translations from GenBank, PDB, SwissProt, PIR and PRF

OMA standalone – Orthologous Matrix, the algorithm which uses evolutionary distances instead of scores, considers distance inference uncertainty, includes many-to-many orthologous relations and accounts for differential gene losses

ORF – Open Reading Frame

OrthoDB - OrthoDB is a comprehensive catalog of orthologs, i.e. genes inherited by extant species from their last common ancestor.

OrthoMCL – Ortholog groups Markov Clustering, the algorithm for grouping proteins into ortholog groups based on their sequence similarity using Markov Clustering

Paired-end sequencing - A major advance in NGS technology occurred with the development of paired-end (PE) sequencing. PE sequencing involves sequencing both ends of the DNA fragments in a sequencing library and aligning the forward and reverse reads as read pairs.

PAM unit - The base unit of time for the PAM matrices is the time required for 1 mutation to occur per 100 amino acids, sometimes called 'a PAM unit' or 'a PAM' of time. This is precisely the duration of mutation assumed by the PAM1 matrix. The constant is used to control the proportion of amino acids that are unchanged.

PANTHER family database - PANTHER is a large collection of protein families that have been subdivided into functionally related subfamilies, using human expertise.

PASA - PASA, acronym for Program to Assemble Spliced Alignments, is a eukaryotic genome annotation tool that exploits spliced alignments of expressed transcript sequences to automatically model gene structures, and to maintain gene structure annotation consistent with the most recently available experimental sequence data. PASA also identifies and classifies all splicing variations supported by the transcript alignments.

PCR – polymerase chain reaction.

PhyloBayes - PhyloBayes is a Bayesian Monte Carlo Markov Chain (MCMC) sampler for phylogenetic reconstruction and molecular dating using protein and nucleic acid alignments.

PhylomeDB - PhylomeDB is a public database for complete catalogs of gene phylogenies (phylomes). It allows users to interactively explore the evolutionary history of genes through the visualization of phylogenetic trees and multiple sequence alignments.

PhyML - PhyML is a phylogeny software based on the maximum-likelihood principle. Early PhyML versions used a fast algorithm performing Nearest Neighbor Interchanges (NNIs) to improve a reasonable starting tree topology. Since the original publication (Guindon and Gascuel, 2003), PhyML has been widely used (>2,300 citations in ISI Web of Science), because of its simplicity and a fair compromise between accuracy and speed. In the meantime research around PhyML has continued, and this article describes the new algorithms and methods implemented in the program (Guindon et al. 2010).

PhymmBL - PhymmBL (rhymes with "thimble"), the hybrid classifier included in this distribution which combines analysis from both Phymm and BLAST, produces even higher accuracy.

QIAamp kit - The QIAamp DNA Mini Kit provides silica-membrane-based nucleic acid purification from tissues, swabs, CSF, blood, body fluids, or washed cells from urine.

Qtrim software - QTrim is executed as a standalone software package for command-line use and integration into sequencing analysis pipelines.

RAxML - A tool for Phylogenetic Analysis and Post-Analysis of Large Phylogenies (Stamatakis 2014).

Reciprocal Best Hit algorithm - Reciprocal Best Hits (RBH) are a common proxy for orthology in comparative genomics.

RefSeq database - RefSeq: NCBI Reference Sequence Database A comprehensive, integrated, non-redundant, well-annotated set of reference sequences including genomic, transcript, and protein.

RepeatMasker - RepeatMasker is a program that screens DNA sequences for interspersed repeats and low complexity DNA sequences.

RNA-seq - RNA-Seq (RNA sequencing), also called whole transcriptome shotgun sequencing (WTSS), uses next-generation sequencing (NGS) to reveal the presence and quantity of RNA in a biological sample at a given moment in time. RNA-Seq is used to analyze the continually changing cellular transcriptome.

samtools - samtools – Utilities for the Sequence Alignment/Map (SAM) format.

sga preqc program - sga comes with a quality control and data exploration module. This module will estimate sequence coverage, per-base error rates and genome size, heterozygosity and repeat content. It is highly recommended to run this module on your data to better understand how difficult the assembly will be. Once you have produced the preqc PDF report, feel free to share it on the sga-users mailing list and ask for advice on how to best proceed with the assembly.

Smith-Waterman algorithm - The Smith–Waterman algorithm performs local sequence alignment; that is, for determining similar regions between two strings of nucleic acid sequences or protein sequences.

SOAPdenovo2 – short-read assembly method.

Tax ID - Taxonomy The Taxonomy Database is a curated classification and nomenclature for all of the organisms in the public sequence databases. This currently represents about 10% of the described species of life on the planet. (https://www.ncbi.nlm.nih.gov/taxonomy)

TransDecoder - TransDecoder identifies candidate coding regions within transcript sequences, such as those generated by de novo RNA-Seq transcript assembly using Trinity, or constructed based on RNA-Seq alignments to the genome using Tophat and Cufflinks. (https://transdecoder.github.io/)

trimAl - trimAl is a tool for the automated removal of spurious sequences or poorly aligned regions from a multiple sequence alignment. (http://trimal.cgenomics.org/)

Trinity - Trinity, developed at the Broad Institute and the [Hebrew University of Jerusalem] (http://www.cs.huji.ac.il), represents a novel method for the efficient and robust de novo reconstruction of transcriptomes from RNA-seq data. Trinity combines three independent software modules: Inchworm, Chrysalis, and Butterfly, applied sequentially to process large volumes of RNA-seq reads. Trinity partitions the sequence data into many individual de Bruijn graphs, each representing the transcriptional complexity at a given gene or locus, and then processes each graph independently to extract full-length splicing isoforms and to tease apart transcripts derived from paralogous genes. (https://github.com/trinityrnaseq/trinityrnaseq/wiki)

USEARCH - The USEARCH algorithm searches a database for high-identity hits to one or more database sequences ("targets"). USEARCH is used by the usearch_global and usearch_local commands and is used as a subroutine by cluster_fast and cluister_smallmem. This algorithm is fundamentally different from the UBLAST algorithm that is designed for low identity local searches. (http://drive5.com/usearch/manual/usearch_algo.html)

WAG substitution matrix – The matrix calculated from a database of globular protein sequences comprising 3,905 amino acid sequences split into 182 protein families (Whelan & Goldman 2001).

wgs database – Whole Genome Shotgun database, contains genome assemblies of incomplete genomes or incomplete chromosomes of prokaryotes or eukaryotes that are generally being sequenced by a whole genome shotgun strategy.